# Automatic Machine Translation Evaluation: A Qualitative Approach

Elisabet Comelles Pujadas

**ADVERTIMENT**. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (**www.tdx.cat**) i a través del Dipòsit Digital de la UB (**diposit.ub.edu**) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

**ADVERTENCIA**. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (**www.tdx.cat**) y a través del Repositorio Digital de la UB (**diposit.ub.edu**) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

**WARNING**. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (**www.tdx.cat**) service and by the UB Digital Repository (**diposit.ub.edu**) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

# Automatic Machine Translation Evaluation: A Qualitative Approach


Elisabet Comelles Pujadas

Tesi presentada per optar al grau de **Doctor en Lingüística** en el programa de doctorat de *Ciència Cognitiva i Llenguatge*, Departament de Lingüística, Universitat de Barcelona,

sota la supervisió de

**Dra. Victoria Arranz Corzana**
Evaluations and Language resources Distribution Agency (ELDA)

**Dra. Irene Castellón Masalles**
Universitat de Barcelona (UB)

Universitat de Barcelona

*A la meva petita-gran família*

# Acknowledgements

This thesis is the result of a six-year research period, during which I have been lucky enough to count on the help of many great people. I owe them a great debt of gratitude and I would like to take this opportunity to thank them all. Without you this thesis would not have been possible.

First of all, I offer my sincere gratitude to my supervisors Dr. Victoria Arranz, from the Evaluations and Language resources Agency, and Dr. Irene Castellón, from the Linguistics Department at the University of Barcelona. Victoria Arranz is the most talented and hard-working woman I have ever had the chance to meet. Her expertise has been a valuable guidance since I started working on research many years ago at the Technical University of Catalonia. Her support, comments and deep knowledge of the area have been crucial during this thesis project and I have been fortunate enough to benefit from them. Gracias, Victoria, por querer emprender esta aventura, por tus sabios consejos y comentarios, por venir siempre al rescate y acompañarme durante todos estos años.

Dr. Castellón is an excellent researcher and professor who I had the fortune to meet when I was studying my degree in Linguistics. Dr. Irene Castellón's broad knowledge of Natural Language Processing and Linguistics has been essential to carry out this research, she has supported me throughout this period with her patience and knowledge, whilst allowing me the room to work in my own way. Gràcies, Irene, per acceptar ser la meva directora, pels teus comentaris i sàvies paraules, per les reunions, els cafès i les teves paraules d'ànim durant aquest viatge.

My thanks are extended to Dr. Jordi Atserias, who has been directly involved in the implementation of VERTa. Dr. Atserias' excellent programming skills and broad knowledge of NLP have been of paramount importance in the process of this thesis; without his help and good advice this thesis would not have come to light. Gràcies, Jordi, per tota la teva implicació i per fer possible la VERTa. Mai podré agrair-te prou tot el temps que ens has dedicat (a la VERTa i a mi), tots els teus savis consells i recomanacions. Gràcies per ajudar-me a arribar fins aquí, sense tu això no hauria estat possible.

I also have to acknowledge all the organisations and persons who have allowed me to use their data in order to carry out my research: Kay Peterson, the National Institute of Standards and Technology (NIST) and the Language Data Consortium for providing the English corpus; the KNOW2 project for providing the Spanish data; Dr. Antoni Oliver, Dr. Salvador Climent and Joaquim Moré for providing extra Spanish references; and Dr. Natalia Judith Laso and Antonio Palacín for volunteering to evaluate the Spanish segments.

During this thesis project I have had the privilege of being a member of the Grial Research Group, where I have widened my knowledge on computational linguistics and Natural Language Processing. This would not have been possible without the help of its members: Dr. Laura Alonso, Dr. Salvador Climent, Dr. Marta Coll, Dr. Ana Fernández, Joaquim Moré, Dr. Antoni Oliver, Dr. Glòria Vázquez, Juan Aparicio, Lara Gil, Marina Lloberes and Jordi Sesé Puértolas. Gràcies a tots pel vostre ajut i en especial al Jordi Sesé per adaptar la sortida del Freeling als requeriments de la VERTa; al Vadó, el Toni i el Quim per la seva col·laboració en les traduccions del castellà; i a la Marina i el Juan pels somriures (i riures), pels moments de distracció i de bon humor tant necessaris al llarg de tot aquest procés.

I am also grateful to all the members of the English Department, where I have been working since I started my research at the University of Barcelona. My deepest acknowledgement to the GreLic group and, especially, to Dr. Isabel Verdaguer for granting me the possibility of collaborating with her research group, and learning about corpus linguistics and discourse analysis. I would also like to thank the other members of this research group as well as those in the teaching innovation group: Dr. Joe Hilferty, Dr. M. Luz Celaya, Dr. Natalia Judith Laso, Dr. Emilia Castaño, Dr. Montse Forcadell, Dr. Sara Feijóo, Dr. Júlia Baron and Aaron Ventura. Moltes gràcies a tots pel vostre suport i tota la vostra ajuda durant aquest llarg camí. És un privilegi treballar i compartir experiències amb tots vosaltres. Isabel, moltes gràcies per tot el teu suport i el teu afecte, així com els ànims que sempre m'has donat. Joe, M. Luz, Montse i Júlia, gràcies també pels vostres ànims i les vostres paraules de suport en els moments difícils. Aaron, gràcies pels riures i els ànims. Ara tu has de ser el següent. Emilia, gracias a ti también por todos los ánimos que me has dado siempre, las charlas sobre la tesis y, por

supuesto, por esas fantásticas fotos de mis Little Prince. Sara, gràcies per la teva amistat i el teu suport durant tots aquests anys, pels riures i les paraules d'ànims en temps difícils. Els dinars al despatx són més agradables si els puc compartir amb tu. Natalia, deixa'm que en aquest paràgraf t'agraeixi tota la teva ajuda a nivell professional. És un privilegi, i sempre ho ha estat, poder treballar amb tu. Ets una persona brillant i tens una capacitat de treball que no deixa de sorprendre'm, n'he d'aprendre tant de tu!

Working at the English Department has also given me the opportunity to get to know other colleagues and friends who I would like to thank for their support during all these years: Dr. Bill Phillips, Ana Marsol, Dr. Marta Ortega, Dr. Rodrigo Andrés, Eva Cerviño, Dr. Clara Escoda, Dr. Imma Miralpeix, Maria Grau and Ursula Wolf. Many thanks to you all.

On a more personal note, I would like to thank all my friends and family, who have been by my side during all these years, who have cheered me up and suffered my ups and downs, especially during these last months of my thesis work. You deserve my most sincere gratitude and I hope life grants me the chance to make up for all this time.

Gràcies a la meva petita-gran família. Sóc afortunada per tenir-vos al meu costat i no vull deixar passar aquesta oportunitat que tinc per agrair-vos-ho. He de començar aquests agraïments esmentant a persones de les que ja he parlat abans: la Victoria, el Jordi i la Natalia. Victoria, nunca podré agradecerte lo bastante toda la ayuda que me has brindado no solo al dirigir esta tesis sino durante todos los años que hace que nos conocemos. Has sido compañera de risas en los buenos tiempos y refugio en tiempos difíciles. Siempre tienes la palabra acertada y sabes qué decir para reconfortarme. Te agradezco mucho que siempre estés ahí. Una altra gran persona a qui li dec molt és la Natalia, una de les persones més bones que conec. Té una capacitat d'estimar infinita i sempre mira pels demés abans que per ella. He tingut molta sort de conèixer-te, amiga, i espero en algun moment poder-te retornar tota la fortalesa i recolzament que m'has ofert quan ho he necessitat. Gràcies per estar sempre al meu costat i ser el meu àngel de la guarda. També vull agrair al Jordi la seva amistat. Batallin, amb tu he compartit moltes coses: la VERTa, viatges, tangos, braves. Sempre has estat present, pacient, escoltant-me i sempre m'has fet riure i m'has animat quan ha calgut. Moltes gràcies!

Part de la meva petita-gran família també són dues estimades grans amigues, la Sandra i la Teresa. Què us puc dir a vosaltres? Fa tants d'anys que ens coneixem que no puc pensar en la meva vida sense vosaltres i sense les vostres famílies. Heu sigut el meu pal de paller quan ho he necessitat i heu tingut una paciència infinita amb mi, en especial aquests últims temps, que degut a la tesi he estat desapareguda. Hem compartit moltes coses plegades, temps alegres i temps tristos, però vosaltres sempre heu estat al meu costat, sempre pendents de mi, ajudant-me a aixecar-me cada cop que he caigut.

Per últim, el nucli més petit de la meva petita-gran família. Per començar el meu pare, una persona lluitadora i treballadora que ho ha fet tot pels seus fills. Gràcies per procurar que rebés una bona educació i per tenir sempre fe en mi. Vull continuar amb el meu germà, una gran persona. Algú que sembla que no hi sigui perquè li agrada passar desapercebut, però que sempre hi és. Sempre pendent de mi, a punt per si el necessites i que és una de les pedres angulars de la meva petita-gran família. No puc oblidar-me de l'Antonio. Has aceptado con resignación quedarte sin fines de semana, vacaciones y puentes ya que yo me había embarcado en esta aventura, y nunca te has quejado, al contrario siempre me has cuidado, mimado y apoyado en los momentos de bajón, incluso cuando afloraba mi carácter gruñón. Por todo esto, gracias! Esta tesis también es un poco tuya.

I per finalitzar, vull agrair a la meva àvia Enriqueta i a la meva mare que sempre em cuidessin i em fessin costat. Com deia Martí i Pol, "*em costa imaginar-te absent per sempre*". Crec que mai podré acostumar-me a la seva absència, però desitjo que siguin on siguin, estiguin contentes del que faig i del que sóc. No podria haver acabat aquesta tesi sense el seu record i la fortalesa que sempre em van transmetre.

# Abstract

The present study addresses the problem of Automatic Evaluation of Machine Translation (MT) from a linguistic perspective. Most of the studies performed in this area focus on quantitative analyses based on correlation coefficients; however, little has been done as regards a more qualitative approach, going beyond correlations and analysing data in detail. This thesis aims at shedding some light on the suitability, influence and combination of linguistic information to evaluate MT output, not restricting our research to the correlation with human judgements but basing it on a qualitative analysis. More precisely, this research intends to emphasize the effectiveness of linguistic analysis in order to identify and test those linguistic features that help in evaluating traditional concepts of adequacy and fluency. In order to perform this research we have focused on MT output in English, with an application to Spanish so as to test the portability of our approach.

The starting point of this work was a linguistic analysis of both MT output and reference segments with the aim of highlighting not only those linguistic errors that an automatic MT evaluation metric must identify, but also those positive linguistic features that must be taken into account, identified and treated as correct linguistic phenomena. Once the linguistic analysis was conducted and in order to confirm our hypotheses and check whether those linguistic phenomena and traits identified in the analysis were helpful to evaluate MT output, we designed and implemented a linguistically-motivated MT metric, VERTa, to evaluate English output. Several experiments were conducted with this first version of VERTa in order to test the suitability of the linguistic features selected and how they should be combined so as to evaluate fluency and adequacy separately. Besides using information provided by correlations as a guide we also performed a detailed analysis of the metric's output every time linguistic features were added and/or combined.

After performing these experiments and checking the suitability of the linguistic information used and how it had to be used and combined, VERTa's parameters were adjusted and an updated and optimised version of the metric was ready to be used. With this updated version and for the sake of comparison, a meta-evaluation of the metric for

adequacy, fluency and MT quality was conducted, as well as a comparison to some of the best-known and widely-used MT metrics, showing that it outperformed them all when adequacy and fluency were assessed.

Finally, we ported our MT metric to Spanish with the aim of studying its portability by checking which linguistic features in our metric would have to be slightly modified, which changes would have to be performed and finally if the metric would be easy to adapt to a new language. Furthermore, this version of VERTa for Spanish was compared to other well-known metrics used to evaluate Spanish, showing that it also outperformed them.

# Resum

Aquesta tesi versa sobre el problema de l'avaluació de la traducció automàtica des d'una perspectiva lingüística. La majoria d'estudis realitzats en aquesta àrea són estudis quantitatius basats en coeficients de correlació, tanmateix, molt poca recerca s'ha centrat en un enfocament més qualitatiu, que vagi més enllà de les correlacions i analitzi les dades detalladament. Aquest treball vol portar llum a la idoneïtat, la influència i la combinació de la informació lingüística necessària per avaluar la sortida de traducció automàtica. En concret, es pretén emfasitzar l'efectivitat de l'anàlisi lingüística per identificar i examinar aquells trets lingüístics que ajudin a avaluar els conceptes tradicionals de fluïdesa i adequació. Per tal de realitzar aquest estudi s'ha treballat amb l'anglès com a llengua d'arribada, tot i que també s'ha tingut en compte el castellà en l'última etapa.

El punt inicial d'aquest treball ha estat una anàlisi lingüística dels segments d'hipòtesi i de referència per tal de trobar tant aquells errors lingüístics que una mètrica automàtica d'avaluació ha de poder detectar, com identificar aquelles característiques lingüístiques que cal tenir en compte i tractar com a fenòmens lingüísticament correctes. Després d'aquesta anàlisi, s'ha dissenyat i implementat una mètrica d'avaluació automàtica, VERTa, que ha d'ajudar a confirmar les hipòtesis formulades i comprovar si els fenòmens i trets lingüístics detectats en l'anàlisi inicial són útils per avaluar text traduït automàticament. Amb aquesta primera versió de la mètrica s'han realitzat una sèrie d'experiments, així com unes anàlisis quantitatives i qualitatives per comprovar la idoneïtat dels trets lingüístics seleccionats i explorar com s'han de combinar per avaluar la fluïdesa i l'adequació per separat.

Després d'aquests experiments i de les anàlisis pertinents, s'han ajustat els paràmetres de la mètrica per tal d'obtenir-ne una nova versió. Aquesta nova versió s'ha utilitzat per realitzar una meta-avaluació de la mètrica, comparant-la amb d'altres mètriques d'avaluació àmpliament conegudes i utilitzades dins de l'àrea. Els resultats obtinguts per la VERTa en relació a l'avaluació de fluïdesa i l'adequació han superat els de la resta de mètriques.

Finalment, s'ha adaptat la mètrica al castellà per tal d'estudiar quines característiques lingüístiques incloses en la mètrica s'havien de retocar, quins canvis calia fer, i si era fàcil adaptar la mètrica a una nova llengua.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1. Introduction

Machine Translation (MT) is one of the most complete tasks within the field of Natural Language Processing (NLP). MT is the automatic translation of one text from a source language into a target language, to put it simply, MT implies using a computer to translate text or speech from one language to another. This is one of the most challenging tasks inside the field of NLP because it implies most types of knowledge that humans possess (i.e. grammar, semantics, knowledge of the world, etc.). According to Basnett (1980), when a person translates a text "*a process of decoding and encoding takes place*" (Basnett 1980:24); in other words, the translation process involves decoding the meaning of the source text and encoding this meaning into the target language, which is a complex cognitive operation. Therefore, decoding a source text means that the translator must understand and analyse the source text entirely, which requires good knowledge of all dimensions of the source language (e.g. lexicon, grammar, semantics) as well as knowledge of the source culture. In addition, the process of encoding also implies the same knowledge of the target language. Reproducing such a complex operation is the challenge of MT. As pointed out by Hutchins and Sommers (1992):

> "The major obstacles to translating by computer are, as they have always been, not computational but linguistic. They are the problems of lexical ambiguity, of syntactic complexity, of vocabulary differences between languages, of elliptical and 'ungrammatical' constructions, of, in brief, extracting the 'meaning' of sentences and texts from analysis of written signs and producing sentences and texts in another set of linguistic symbols with an equivalent meaning." (Hutchins and Sommers 1992:2)

In other words, the challenge or complexity behind MT lies in programming a computer so that it follows the same process as a human translator does: understanding a text as a person does, with all the knowledge that it implies, and being capable of creating a new text in the target language, with all the knowledge of the target language and target culture that goes with it.

Directly linked to MT there is a subtask, Machine Translation Evaluation, which is intended to check (or evaluate) the quality of the automatic translation produced. As pointed out by Hutchins and Sommers (1992), there are several types of evaluation which can be performed at different stages:

- evaluation performed during the development of a system (e.g. to check the effects of any changes in the system);
- evaluation once the system has been developed before offering it to a potential user (e.g. to check the robustness and computational efficiency of a programme, to test the integration of a system in a particular computer environment);
- evaluation of the system by its potential buyers and users (e.g. to check the quantity and kind of human input necessary to produce acceptable translations);
- evaluation of the system by the final recipients of translations (e.g. to compare human translation with machine translation in terms of speed and quality).

In most of these stages there is one common point, the linguistic quality of the MT output. In order to evaluate this MT output one can focus on the assessment of the MT quality or on error analysis. Whereas the former deals with aspects such as assessing the accuracy or fidelity in translating the meaning of the source sentence or assessing if the target sentence can be understood, the latter focuses on identifying and classifying errors made by the MT system.

Both types of evaluations were initially performed by human evaluators. This has the advantage that MT developers are provided with a wide range of assessments regarding partial aspects of MT quality (ALPAC report 1966; White et al. 1994; Snover et al. 2006; Lo and Wu 2011). In addition, human evaluators possess all that knowledge that MT systems try to emulate. On the other hand, performing this type of evaluation is very expensive, time-consuming and subjective – sometimes the inter- and/or intra-annotator agreement is rather low (Turian et al. 2003; Ye et al. 2007; Callison-Burch et al. 2012). As a reaction to these drawbacks and since MT developers required fast and reliable MT evaluations, the MT community started developing and using automatic MT evaluation measures, the framework of this thesis.

Automatic MT evaluation metrics are supposed to be faster, cheaper and more objective than human evaluation. Actually, the use of this type of evaluation has been widely extended among MT developers because they can carry out fast evaluations of their MT systems and immediately use the results obtained to improve them. This is the main reason why in the last decade a wide range of MT metrics has been developed. Most of them work as similarity measures and use reference translations to compare them to the MT output or hypothesis. Among these there are BLEU (Papineni et al. 2001), NIST (Doddington 2002), METEOR (Banerjee and Lavie 2005), SMT and HWCM (Liu and Gildea 2005), TER (Snover et al. 2006), SR (Giménez 2008a), MEANT (Lo and Wu 2012) or DiscoTK (Joty et al. 2014), just to name some of them. Other metrics are aimed at estimating MT Quality, in other words, predicting the quality of MT output when reference translations are not available, such as those metrics proposed by Specia (2009/2010/2011).

From those metrics using similarity measures, some do not use linguistic information at all, such as BLEU and NIST among others; some use information at lexical level (e.g. synonyms, stemming, paraphrasing) such as METEOR, M-TER and M-BLEU (Agarwal and Lavie 2008), TERp (Snover et al. 2009) or SPEDE (Wang and Manning 2012); some use morphological information (e.g. information about suffixes, roots, prefixes) such as AMBER (Chen et al. 2012) and INFER (Popoviç 2012); some use information regarding morphosyntax (e.g. Part-of-Speech (PoS) tags, constituents, dependency relations) such as SMT and HWCM (Liu and Gildea 2005), Owczarzak et al. (2007a/b), SP, CP and DP metrics (Giménez 2008a), DepRef (Wu et al. 2013); some make use of information related to semantics, such as SR and DR metrics (Giménez 2008a), SAGAN-STS (Castillo and Estrella 2012), MEANT (Lo et al. 2012) and UMEANT (Lo and Wu 2013). Most of the above mentioned metrics evaluate partial aspects of MT output (e.g. vocabulary, syntax, semantics); however, in the last years MT metrics have been more oriented to evaluating MT quality in general and MT researchers have struggled to find the best way to combine different types of MT metrics either by using machine learning techniques (Albrecht and Hwa 2007a and 2007b; Yang et al. 2011; Gautam and Bhattacharyya 2014; Joty et al. 2014) or trying more simple approaches such as MAXSIM (Chang and Ng 2008), ULC (Giménez and Márquez 2010b), IPA and STOUT (González et al. 2014).

The above mentioned metrics range from very simple metrics, usually aiming at partial aspects of quality, to highly sophisticated ones, using a large amount of information and machine learning techniques. It must also be highlighted that the performance of these metrics depends on how well they correlate with human judgements and they are developed and improved taking into account these correlations.

Giménez and Márquez (2010b) reported that linguistic information and especially their combination of linguistic features correlated well with human judgements in several evaluation campaigns. However, little qualitative analysis on the use and influence of linguistic features, regardless of how well or badly they correlate with human judgements, has been performed. We consider that this qualitative analysis is also appropriate since, although correlation with human judgements is the standard method to evaluate the performance of a metric, it is highly dependent on the degree of intra-/inter-annotator agreement (Turian et al. 2003; Callison-Burch et al. 2012). Furthermore, when using more sophisticated metrics that combine linguistic information at different levels, such as those reported above, it is hard to interpret their score since this type of metrics uses such highly heterogenous types of linguistic features that it is difficult to know to what extent and how each linguistic feature is contributing to the evaluation of MT output.

### 1.1 Aim of this Thesis and Main Hypotheses

The research presented in this thesis aims at shedding some light on the suitability, influence and combination of linguistic information to evaluate MT output, not restricting our research to the correlation with human judgements but especially basing it on a qualitative analysis. More precisely, this research intends to emphasize the effectiveness of linguistic analysis in order to identify and test those linguistic features that help in evaluating traditional concepts of adequacy and fluency. We move away from evaluating MT quality in general, since we consider that from a linguistic point of view, a divide and conquer strategy will be more effective and appropriate to test the validity and especially the influence of the linguistic features we intend to use. Therefore, this study is mainly based on the analysis of data and the performance of experiments from a linguistic perspective. In order to perform this research we have

focused on MT output in English, although MT output in Spanish has also been taken into consideration.

From this general aim, some hypotheses and sub-hypotheses have been formulated:

**Hypothesis 1. A linguistic analysis can help to clarify what linguistic features should be used and how they should be combined to evaluate MT output**. This hypothesis addresses the issue whether a linguistic analysis can help to identify the most appropriate linguistic information and how it should be used to evaluate MT output in English. In this sense, we will explore if the correlation with human judgements, the standard process used in the meta-evaluation of metrics, coincides with a more linguistic approach or there are some discrepancies and if the combination of both methodologies is useful.

**Hypothesis 2**: Last evaluation campaigns (e.g. Workshops on Statistical Machine Translation WMT08 – WMT14) have reported that those MT metrics using linguistic information correlated better with human judgements on MT quality. Thus, MT metrics improve when adding linguistic information. However, when evaluating MT quality in general, and especially when it is assessed by means of a ranking approach, it is difficult to identify how linguistic knowledge is helping in the evaluation. We think that **addressing fluency and adequacy evaluations separately would help to easily identify the use and suitability of linguistic features. Therefore, linguistic features would be more or less appropriate depending on the type of evaluation, either adequacy or fluency**. This can be broken down into the following specific points:

i. **Organising linguistic information at different levels and aiming at different tasks might help to detect MT errors, which might be especially useful to improve knowledge-based MT systems**.

ii. **Lexical semantics helps to evaluate adequacy. Most of the linguistically-enhanced metrics use synonyms, but we think that other type of lexical semantic relations, such as hypernyms and hyponyms, might also help to evaluate adequacy**.

iii. Most of the MT metrics using syntactic information (i.e. constituent and dependency analyses), have reported to correlate well with judgements on fluency (Liu and Gildea, 2005; Owczarzak et al. 2007a/b). Actually, Lo and Wu (2010) stated the following:

> "Unlike the widely-used lexical and n-gram based or syntactic based MT evaluation metrics which are fluency oriented, our results show that using semantic role labels to evaluate the utility of MT output achieve higher correlation with human judgments on adequacy". (Lo and Wu 2010)

However, we think that **depending on how syntactic information is used it can help to evaluate both adequacy and fluency**.

iv. Information regarding Semantic Roles (SR) and Named Entities (NE) has been used to evaluate adequacy (Lo and Wu 2010; Lo et al. 2012). We think that **other semantic information such as Sentiment analysis, NE linking and identification of Time Expressions can also help to evaluate adequacy**.

**Hypothesis 3**: Giménez and Márquez (2010b), as well as the top metrics in WMT14 (Joty et al., 2014; Gautam and Bhattacharyya, 2014), showed that the combination of linguistic information at different levels helps in the evaluation of MT quality. However, they focus on MT quality in general and use a wide range of metrics; thus making it difficult to know how each of the metrics contributes to the evaluation. We think that **studying different evaluation tasks might not be only useful to identify which linguistic features are more or less appropriate depending on the type of evaluation (adequacy or fluency) but also how they should be combined**.

i. **In order to evaluate the fluency of a segment, that information aimed at checking the grammaticality of a sentence seems to be the most convenient: morphosyntactic information (i.e. lemma, PoS), word order and dependency relations.**

ii. **In order to evaluate the adequacy of a segment, that information related to both lexical and dependency relations seems to be the most relevant one**. According to the principle of compositionality (Frege's Principle) "the meaning

of a whole is a function of the meaning of the parts and of the way they are syntactically combined". Thus, **the interaction between lexical semantics and dependency relations should account for the meaning of the sentence**.

**Hypothesis 4**: Depending on the source and target language, the type of linguistic features used and how they are combined might vary. **Thus, porting a linguistically-enhanced MT metric to a new language may involve studying the main key features of that language and reflecting them on how linguistic features are used in the metric**. To confirm this hypothesis we aim at porting an MT metric from English into Spanish to evaluate adequacy considering the following:

i. **Information on PoS might be disregarded when evaluating adequacy in English, but it might be useful when addressing Spanish.**

ii. **Word order might have a stronger influence when evaluating English than when evaluating Spanish, since word order in Spanish is more flexible than in English.**

This section has set the aim of this thesis and our main hypotheses. Next, the organization of this thesis is described.

### 1.2 Thesis Structure

This thesis consists of 7 chapters, namely:

- Chapter 1. Introduction
- Chapter 2. State of the Art
- Chapter 3. Methodology
- Chapter 4. Linguistic Analysis of the Data
- Chapter 5. VERTa. Metric Description
- Chapter 6. Experiments on Adequacy
- Chapter 7. Experiments on Fluency
- Chapter 8. Meta-evaluation of VERTa
- Chapter 9. Porting VERTa to Evaluate Adequacy for Spanish
- Chapter 10. Main Contributions and Future Work

The present chapter is a brief introduction to the research performed in this thesis. The framework of our work is briefly overviewed and the main aim of our research is stated (section 1.1): exploring the suitability, influence and combination of linguistic information to evaluate MT output from a linguistic point of view. This main aim can be broken down into 4 hypothesis: firstly, the validity of a linguistic analysis to clarify what linguistic information should be used and how it should be combined to evaluate MT output; secondly, the most relevant linguistic features to evaluate adequacy and fluency in English; thirdly, the variability of these features and their combination depending on the type of evaluation (i.e. adequacy or fluency); and finally, the variability of the linguistic features and their combination depending on the language evaluated (i.e. English vs. Spanish). Finally, section 1.2. provides an outline of the main chapters of the present study.

Chapter 2, "State of the Art", focuses on MT evaluation and its different types. It especially emphasises automatic MT metrics and the linguistic information they use, since this is part of the foundation of the present study. Section 2.2 covers *non-automatic evaluations* and presents two different approaches: context-based evaluation and human evaluation of MT quality. Section 2.3 narrows the focus down to *automatic evaluation* and distinguishes between automatic evaluation without reference translations (2.3.1) and the heart of our research, automatic evaluation with reference translations (2.3.2). This last subsection offers a description of the most well-known and widely-used MT evaluation metrics and divides them according to the type of linguistic information that they use: no linguistic information (2.3.2.1), lightweight linguistic information (2.3.2.2), heavyweight linguistic knowledge (2.3.2.3) and metrics combination (2.3.2.4). Finally, section 2.4 is a brief summary of the metrics described and their pros and cons.

Chapter 3, "Methodology", describes the methodology used to conduct this research. As stated in this chapter we have followed an empirical approach, thus grounding our study on the linguistic analysis of data and the experiments conducted. Section 3.1 describes the steps followed to conduct this research. Section 3.2. details the data, resources and tools used in this present work. Finally, 3.2 is a summary of the main points in this chapter.

Chapter 4, “Linguistic Analysis” describes the linguistic analysis carried out with part of our development corpora for English and for Spanish. This analysis is aimed at identifying those linguistic phenomena that must be taken into account when evaluating MT output with reference translations either because they are translation errors or positive linguistic features that must be considered. The linguistic phenomena described are organised into different linguistic levels. Thus, section 4.1 is devoted to format and orthography; section 4.2 covers linguistic phenomena at lexical level, section 4.3 deals with morphology; section 4.4 describes syntactic phenomena; and section 4.5 presents phenomena at semantic level. Finally, section 4.6 is a summary of the most salient linguistic phenomena.

Chapter 5, “VERTa: an MT metric”, describes the MT metric that we have developed taking into account the most salient linguistic phenomena described in Chapter 4 and which will serve us as a tool to perform the experiments to explore and check the suitability of linguistic information for MT evaluation. Section 5.1 proposes the classification of the linguistic phenomena that we will cover into different levels. Section 5.2 describes the architecture of the metric, how the metric works and its organisation into different modules corresponding to the organisation of the linguistic phenomena. Finally, section 5.3 offers a brief summary of the contents of this chapter.

Chapter 6, “Experiments on Adequacy”, is the first chapter aimed at describing the experiments performed and discussing the results obtained. This chapter focuses on those experiments carried out to explore the suitability of those linguistic features included in VERTa to evaluate adequacy at segment level. The results obtained are discussed taking correlations with human judgements as a starting point but with a focus on providing a more qualitative analysis using a linguistic approach. Section 6.1 describes the data used to perform the experiments; section 6.2 describes experiments performed with the Lexical Module; section 6.3 is aimed at the Morphological Module; section 6.4 describes experiments performed with the N-gram Module; section 6.5 is aimed at the Dependency Module; section 6.6 tests the suitability of the Semantic Module; section 6.7 is devoted to the combination of linguistic features and modules to

evaluate adequacy trying to achieve the best possible combination; and finally, section 6.8 draws some conclusions on the experiments performed.

Chapter 7, “Experiments on Fluency”, is aimed at presenting the experiments and discussing the results obtained when testing linguistic information to evaluate the fluency of a segment. Section 7.1 explores the use of the Lexical Module; section 7.2 presents experiments on the Morphological Module; section 7.3 deals with the N-gram Module; section 7.4 presents experiments on the Dependency Module; section 5 tests the Semantic Module; section 7.6 deals with the use of a Language Model (LM); section 7.7 checks how linguistic features and modules should be combined, as well as their influence when evaluating the fluency of a segment and aiming to obtain the optimal combination; and finally, conclusions on the findings obtained in the experiments are drawn in section 7.8.

Chapter 8, “Meta-evaluation of VERTa”, is aimed at confirming VERTa, which was initially developed as a tool to perform our study and our experiments, as an efficient MT metric. To this aim a meta-evaluation of the metric has been carried out and VERTa has been compared to other well-known metrics. Section 8.1 presents the meta-evaluation to test adequacy: the metrics against which VERTa has been compared are described, then results of the meta-evaluation are presented and discussed, providing both quantitative and qualitative analyses. Section 8.2. is aimed at the meta-evaluation to test fluency: first metrics to which VERTa is compared are described, then results obtained are presented and discussed by means of a quantitative and a qualitative analysis. Section 8.3 deals with the participation of VERTa in the WMT14 Metrics Shared Task: first some preliminary experiments performed before participating are presented, then VERTa’s results and participation in the WMT14 Metrics Shared Task is discussed. Finally, conclusions regarding the meta-evaluation of VERTa are drawn (section 8.4).

Chapter 9, “Porting VERTa to Spanish”, presents a first approach to exploring the linguistic features that have to be modified and how they have to be modified when porting the English version of VERTa to Spanish in order to evaluate the adequacy of a segment. Section 9.1 presents the goal and data used to perform this experiment, section 9.2 is aimed at the experiments performed to check the suitability of the linguistic

information in Spanish and how to combine the modules in VERTa. Section 9.3 compares the English version of VERTa to evaluate adequacy with the Spanish version to highlight the points in common and where both versions differ. Then, for the sake of comparison, VERTa's performance is compared to that of other well-known metrics in section 9.4. Finally, section 9.5 sums up the main findings of this chapter.

Chapter 10, "Conclusions and Future Work". This chapter concludes the present thesis and highlights its main contributions. Section 10.1 revisits the initial hypothesis and checks whether they have been confirmed by the present study. Section 10.2 points out the main contributions of this thesis. Then, section 10.3 deals with future work and new research lines.

Finally, the bibliographical references used and 3 appendices conclude this thesis.

Next, the chapter "State of the Art" opens the door to this thesis, setting up the framework and previous research on which this thesis has been grounded.

# Chapter 2. State of the Art

Machine Translation (MT) is directly linked to MT Evaluation since it plays a key role in the MT development cycle in order to improve already existing MT systems as well as to develop new MT strategies. In addition, MT evaluation can also be crucial for MT users, since it may help them find the MT system that best fulfills their needs. As stated in Chapter 1, MT is a very complex task since it implies understanding and producing natural language. Similarly, evaluating MT output also implies performing a complex process: understanding a sentence and decide whether it has been correctly translated Throughout the history of MT Evaluation, several approaches and methodologies have been proposed, developed and used. This chapter aims at providing an overview of MT evaluation, focusing on its different types, as well as discussing their weak and strong points. MT evaluation has been classified into two main types, non-automatic evaluation (section 2.1) and its subtypes (section 2.1.1 and 2.1.2) and automatic evaluation (section 2.2). Special emphasis is placed on the latter and its two main approaches: MT evaluation without reference translations (section 2.2.1) and MT evaluation using reference translations (section 2.2.2). Since MT evaluation using references is the framework for the research presented in this thesis, this type of automatic evaluation will be analysed in detail, presenting the different MT metrics available nowadays and the information they use. Finally, section 2.3 draws some conclusions.

## 2.1 Non-automatic MT Evaluation

Non-automatic MT evaluation, in other words, evaluation performed by people, is usually divided into context-based evaluation and MT-quality evaluation, hereafter referred to as human evaluation. This section starts by describing context-based evaluation briefly (section 2.1.1), a more user-oriented evaluation, and later, human evaluation (section 2.1.2), a type of evaluation focused on the quality of MT output, emphasizing its different types and approaches.

### 2.1.1 Context-based Evaluation

Context-based evaluation stresses the idea that potential users of MT technology should first evaluate the suitability of both the MT Technology (e.g. statistical MT, rule-based MT, hybrid MT) and the MT system for their specific purpose. Church and Hovy (1993) started analyzing this approach, and their work was continued by the Evaluation Working Group of the ISLE Project (1999-2002). They developed FEMTI, an MT evaluation framework (Hovy et al. 2002), which helps MT users to evaluate MT systems according to a wide range of characteristics and quality aspects such as functionality, reliability, efficiency, maintainability, portability, cost, etc. This framework[1] is available to the user and has been developed within the work of the ISSCO team (Hovy et al. 2002; Estrella et al. 2005; Popescu-Belis et al. 2006). FEMTI is rather complex since it considers a large amount of parameters and quality aspects.

### 2.1.2 Human Evaluation

Human or manual evaluation is a quality-oriented evaluation, since the quality of the translations generated by an MT system is evaluated. In contrast to context-based evaluation, this type of evaluation is more suitable for MT developers since they are allowed to measure the quality of the output produced by their systems. However, this type of evaluation is time-consuming, subjective and rather expensive. Several methodologies and approaches to human evaluation have been suggested, some of the most well-known being the following:

- **Fidelity and Intelligibility.** These two measures were proposed in the ALPAC report (1966). Fidelity (or accuracy) is aimed at measuring how much information the translated sentence retained compared to the original (on a scale of 0-9), whereas intelligibility is aimed at measuring how "understandable" the automatic translation is (on a scale of 1-9). Inteligibility was measured without referring to the original sentence, while fidelity was measured indirectly. Raters were first asked to gather whatever they could from the translated sentence and then they were asked to evaluate the original sentence on its informativeness, taking into account what they

[1] http://www.ilc.cnr.it/EAGLES/isle/ISLE_Home_Page.htm

had understood from the translated sentence. Therefore, if the original sentence was rated as "highly informative" this meant that the translated sentence lacked fidelity. These measures proved useful to distinguish human translations from automatic translations.

- **Adequacy and Fluency.** Another method proposed by ARPA[2] consists in a group of human evaluators who are presented a translation segment and have to rate it according to two parameters: adequacy and fluency on a scale of 1-5 (see Table 1). Adequacy is related to the content and semantics, and refers to the degree to which the information present in the input sentence is also communicated in the output sentence. Fluency is related to syntax and refers to the degree the output sentence is well-formed according to the rules of the target language.

| Scores | Adequacy | Fluency |
|---|---|---|
| 5 | All information | Flawless English |
| 4 | Most | Good |
| 3 | Much | Non-native |
| 2 | Little | Disfluent |
| 1 | None | Incomprehensible |

**Table 1** Interpretation of Adequacy and Fluency scores

Although this method has been used for a long time, some evaluators claimed that *"The manual evaluation of scoring translation on a graded scale from 1–5 seems to be very hard to perform"* (Koehn and Monz 2006) and others have also pointed out the low agreement between human judges (Callison-Burch et al. 2012).

- **Ranking of full sentences.** This measure has been used by WMT since 2008, works at sentence level and its aim is to compare up to five MT output sentences from different systems and rank them from best to worst (ties allowed) on whatever criteria the annotator thinks appropriate. Although this evaluation method seems to be faster and reach a better agreement between annotators than the absolute fluency and adequacy method, it is "*still far from satisfactory*" according to Bojar and Wu (2012). Several discrepancies have been observed in the interpretation of the

[2] http://www.darpa.mil/default.aspx

rankings, partly due to the difficulties in ranking very long sentences and also due to the technicalities of the calculation, e.g. including or disregarding ties leads to different ranking of the systems. [see Bojar et al. (2011) through Bojar and Wu (2012)]. In addition, this method evaluates quality in general, thus making it more difficult to know why one system is better than another.

- **Post-Edit Time.** This measure consists in calculating the time required by the post-editor to transform the output sentence of an MT system into a valid translation.

- **HTER** (Human-targeted Translation Edit Rate) (Snover et al. 2006). This method requires human evaluators who are fluent in the target language to generate a new targeted reference. Human annotators are given the MT output and one or more predetermined references and they are asked to generate a new targeted reference by editing the MT output sentence so that it has the same meaning as the reference and is understandable. Later, the minimum TER (Translation Edit Rate) is calculated using the new targeted sentence as a new human reference, in other words, a program compares the unedited MT output sentence to the human-edited one and to the reference sentence and finds the minimum number of edits performed. This method implies that the human evaluator must understand the meaning of the reference translation and also propose the minimum number of edit changes to the MT output, with the aim that the MT output expresses the same meaning as the reference translation. Therefore, HTER requires trained people and it results quite expensive.

- **HMEANT** (Lo and Wu 2011). This method focuses on the predicate-argument structure of the sentence and uses Semantic Roles (SR) information to assess the *utility* of MT output. Evaluators are asked to check *«if they recognize "who did what, to whom, when, where and why"(Pradhan et al. 2004) from the MT outputs and whether the respective role fillers convey the same meaning as in the reference translation»* (Bojar and Wu 2012). Therefore, this method consists in identifying the semantic frames and roles (SRL) in the hypothesis and reference translations,

aligning these frames and role fillers and finally calculating the precision[3] and recall[4] across all frames in the sentence. The authors claim that HMEANT correlates with human judgements on adequacy as well as HTER but it does not require trained evaluators, thus involving a low labour cost. Although this seems to be a very interesting metric, it is mainly focused on adequacy so it might not be useful for a user whose aim is to check how fluent the MT output is (i.e. post-editors).

Although human evaluations are very informative, they present several weak points: they are expensive, time-consuming and subjective. Firstly, issuing judgements on quality implies hard work on the evaluator's side, which results in current evaluation campaigns and shared tasks only producing human judgements for a subset of sentences and systems. Secondly, as human evaluation is such a labor-intensive task, human judges spend an important amount of time evaluating MT outputs, thus turning human evaluation into a time-consuming activity. Finally, as it is a task performed by people, assessments may vary from one evaluator to another, which results into low agreement between judges (Callison-Burch et al. 2012). Moreover, the guidelines provided to the judges may differ between evaluation campaigns. Last but not least, we must also consider other factors such as the knowledge of the language, which may differ from one judge to another. As a result, evaluation campaigns have moved towards reducing human assessment and increasing the use of automatic metrics by means of shared tasks, for instance 2004 and 2005 NIST Evaluation Campaigns, 2006 TC-STAR Evaluation Campaign, 2008-2010 MetricsMATR, WMT 2008-2014, where the objective was to automatise the process as much as possible.

### 2.2 Automatic Evaluation

In opposition to manual evaluations, automatic evaluations are not so expensive, and they are faster and more objective (if the same metric is run on one output sentence the result obtained will always be the same, whereas the judgement of a human evaluator can change) and updatable. The most outstanding characteristic of this type of

[3] The proportion of elements in the hypothesis translation that can be found in the reference translation.
[4] The proportion of elements in the reference translation that can be found in the hypothesis translation.

evaluation is that it is much faster than human evaluation and this allows developers to use the results obtained immediately in the system development cycle. On the other hand, automatic evaluation is not the panacea for evaluating MT output because most of the automatic evaluations are partial and often devoted to shallow aspects of quality (ngram-based metrics, lexical-based metrics, syntax-based metrics or semantic-based metrics). Moreover, they depend on the availability of a set of reference translations, the development of which is also time-consuming and expensive. To overcome this latter weakness, several researchers have approached MT evaluation disregarding reference translations.

In the following we provide an overview of automatic evaluation without references (section 2.2.1) and later a detailed presentation and analysis of automatic evaluation using reference translations (section 2.2.2).

### 2.2.1 Automatic Evaluation without Reference Translations

As presented above, traditional MT metrics show a main drawback, they depend on reference translations, whose development is expensive and time-consuming. In opposition, some researchers have started to work on the use of automatic MT metrics without reference translations. The most relevant systems are described below.

- Quirk (2004) followed an approach which predicted MT quality and fluency at sentence level without the use of reference translations, but relying on human assessments. Several features were gathered by means of their hybrid machine translation system MSR-MT. The first group of features refers to the source sentence and how difficult it is to parse (e.g. size of the input sentence). The second group addresses features about the translation process itself (e.g. number and average size of the learned mappings). Finally, features about the proportions of words and substrings covered by the learning corpus. Next, a variety of supervised machine learning algorithms were applied. Results obtained were satisfactory, although this method was only tested on the MT output of a single system, and therefore, it is difficult to know how well this would generalize.

- Gamon et al. (2005) aimed at evaluating MT quality and fluency without reference translations. They combined language model perplexity scores with class probabilities from a machine-learned classifier which used linguistic features (i.e. trigrams of Part-of-Speech (PoS), syntactic tree nodes and their labels) and had been trained to distinguish human translations from machine translations.

- Hamon and Rajman (2006) proposed a new metric called X-Score, a fluency-oriented metric, which relies exclusively on the syntax of the target document, without using reference translations. This metric is based on morpho-syntactic categories or syntactic relations and a fluency score obtained from a training corpus.

- Albrecht and Hwa (2007b) proposed a model that tries to capture features on fluency and adequacy. In order to identify those features related to adequacy the input sentence is compared to pseudo-references, which are sentences from other MT systems; whereas, traits related to fluency are obtained by comparing the input to target language resources such as treebanks and large text corpora.

- Specia et al. (2009/2010) aimed at predicting the quality of MT output when reference translations are not available by means of addressing the problem as a regression task. They trained algorithms to predict different types of sentence-level scores and used a feature selection strategy in order to exploit a large number of features which included those extracted from the input and output sentences and those that depend on the translation process of a given MT system.

- Parton et al. (2011) presented MTeRater and MTeRater-Plus in the 6$^{th}$ Workshop on Statistical Machine Translation. Both metrics are machine-learned metrics that use features from e-rater®, an essay scoring engine that assesses writing proficiency. MTeRater addresses fluency issues and MTeRater-Plus accounts for adequacy by combining MTeRater with other MT evaluation metrics and heuristics.

- Avramidis et al. (2011) proposed a method that uses a statistical classifier trained upon existing human rankings, using several features (sentence length, ngrams information, parsing information and shallow grammatical matches) derived from analysis of the source and target sentences.

- Specia et al. (2011) proposed an approach to predict the adequacy of MT output at sentence level. They used a machine learning algorithm trained on translations which had been previously assessed and several quality indicators on adequacy.

- Lo et al. (2014) presented a new metric XMEANT, a new cross-lingual version of MEANT, a semantic-frame based MT evaluation (see section 2.2.2.3 for further details), which does not use reference translations.

These approaches have been highly valued in order to predict the quality of machine translation when no reference translations are available. In fact, in the last years, several shared tasks on Quality Estimation (QE) have been held (WMT12-WMT14). However, when reference translations are available, traditional automatic MT metrics are still preferred. As Specia et al. (2010) stated:

> "The QE metric is not meant to replace evaluation metrics, but instead provide a way to assess quality when reference translations are not available".

Thus, according to Specia et al. (2010), QE metrics do not pretend to substitute reference-based MT metrics, but complement them.

Next, we give an overview of the most well-known approaches to automatic MT evaluation with reference translations.

### 2.2.2 Automatic Evaluation with Reference Translations

As seen so far, MT evaluation is essential in the development cycle of MT and automatic metrics are preferred to human evaluation because, and even though some of them make use of references that are also costly, it is undeniable that automatic metrics are faster and more objective than human evaluation. This is the reason why in the last

decade a wide range of automatic MT metrics using reference translations has been developed. This section presents the most relevant ones.

Traditionally, automatic MT metrics are organized according to the way they calculate their final score:

a. Edit-Distance Measures: These measures are based on the number of changes which need to be applied to the output of the MT system to transform it into a reference translation (Tillman et al. 1997, Nieβen et al. 2000; Snover et al. 2006 and 2009; Leusch and Ney 2009; Wang and Manning 2012).

b. Precision-Oriented Measures: These metrics are based on lexical precision, that is, the proportion of lexical units in the automatic translation covered by the reference translations (Papineni et al. 2001; Doddington 2002; Leusch and Ney 2008).

c. Recall-Oriented Measures: These metrics are based on lexical recall, i.e., the proportion of lexical units in the reference translations covered by the automatic translation (Lin and Och, 2004; Leusch et al. 2006/2009).

d. Measures Balancing Precision and Recall: These metrics are a combination of precision and recall (Melamed et al. 2003; Turian et al. 2003; Banerjee and Lavie, 2005; Chang and Ng, 2008; Liu et al. 2010).

However, since the purpose of this research is to study the influence and use of linguistic information in MT evaluation, a classification based on the type of linguistic information used has been preferred[5]. Therefore, in this section MT metrics are classified as follows:

e. No-linguistic Knowledge Metrics (section 2.2.2.1). Metrics that do not use any linguistic knowledge.

[5] For the sake of clarity, a table with the most relevant linguistically-motivated metrics is provided in Appendix C.

f. Lightweight Linguistic Knowledge Metrics (section 2.2.2.2). Metrics that use linguistic knowledge at low level, such as information on lemmas, stemming, synonyms and paraphrasing.

g. Heavyweight Linguistic Knowledge Metrics (section 2.2.2.3). Metrics that use richer linguistic knowledge by means of NLP tools (e.g. PoS, constituent parsing, dependency parsing, SR labeling, Textual Entailment (TE), etc.).

h. Combination of metrics (section 2.2.2.4). How metrics covering different aspects of MT quality and linguistic information at different levels are combined.

#### 2.2.2.1 No-linguistic Knowledge Metrics

As pointed out above, these metrics do not use any kind of linguistic knowledge or linguistic features to evaluate MT output. All of them are based on lexical similarities (also called *n*-gram based metrics). This approach uses a set of reference translations and seeks the lexical similarity (*n*-gram matching) between the MT output and these reference translations with no use of linguistic features. The main differences among them are how the lexical similarities are calculated.

- *Word Error Rate* (WER) (Nieβen et al. 2000). An edit-distance measure that derives from Levenshtein distance (Levenshtein 1966), although applied at word level instead of phoneme level, and which takes into account the number of deletions, insertions and substitutions that must be performed in order to obtain a valid translation. Although WER was widely used during a period of time, it neither allows reordering of words, nor provides any information on the nature of errors.

- *Position Independent Word Error Rate* (PER) (Tillman et al. 1997). Also an edit-distance measure similar to WER but in opposition to this, it does not take into account the order of words in the sentence.

- *Bilingual Evaluation Understudy* (BLEU) (Papineni et al. 2001). This is with no doubt the most frequently used metric in the MT field, since it is fast and easy to use. BLEU is a precision-oriented measure which compares n-grams of the hypothesis translation with n-grams of the reference translation and

counts the number of matches. These matches are position independent and can go up to length 4. The more the matches obtained, the better a translation is. In addition, BLEU also applies a brevity penalty which penalizes candidates that are shorter than their reference, and which modifies the overall BLEU score.

- The NIST metric, developed by the *National Institute of Standards and Technology* (NIST) (Doddington 2002). NIST is an improved version of BLEU that weighs more heavily those n-grams that are more informative. That is to say, NIST weighs more heavily those n-grams that occur less frequently because their informative value is higher. In addition, NIST takes into account n-grams up to length 5. It also presents some changes in the calculation of the brevity penalty, since small variations in the translation length do not affect much the overall score.

- *General Text Matcher* (GTM) (Melamed et al. 2003; Turian et al. 2003). This metric balances precision and recall, it is based on the F-measure[6]; and allows for different weights depending on the length of the n-gram (longer n-grams are assigned a higher weight than shorter n-grams).

- WNM (Weighted N-gram Model) (Babych and Hartley 2004). This measure is a variant of BLEU that includes statistical weights which capture n-grams' degree of salience estimated out from a monolingual corpus.

- *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) (Lin and Och 2004). This measure calculates recall over n-grams. It can evaluate separately 1, 2, 3 and 4 n-grams. Besides no length penalty is applied.

- *Translation Edit Rate* (TER) (Snover et al. 2006). TER, the automatic measure related to human evaluation metric HTER (see section 2.1.2), was introduced by the GALE (Global Autonomous Language Exploitation) research program (Olive 2005). This metric calculates the minimum number of edits that a post-editor has to make to a system output so that it exactly

[6] The harmonic mean of Precision and Recall.

matches one of the reference translations. All edits have equal cost and possible edits include deletion, insertion, substitutions of single words, and in opposition to WER, it also allows for shifts of word sequences.

- BLEUSP (Leusch and Ney 2008). This MT metric is an improved version of BLEU which uses a smoothed n-gram geometric mean in order to combine n-gram precisions. Segment boundary markers are also used so as to increase the weight of words near the segment boundaries in the BLEU score.

The adoption of these metrics implied a great advance in MT research. They have been widely accepted and used by the SMT research community, and particularly BLEU, has been adopted as the "de facto" standard evaluation method, mainly because it is easy to use, quick and inexpensive. However, it has also received many criticisms particularly by those not developing SMT systems. While it performs quite well on assessing fluency, Callison-Burch et al. (2006) and Koehn and Monz (2006) reported cases where BLEU strongly disagreed with human judgement on translation quality, especially when statistically-based MT systems are compared to rule-based systems BLEU shows a tendency to assign better scores to the former systems. This can also be extended to n-gram-based metrics, since they work in a similar way to statistical MT systems. Thus, they favour those hypotheses that contain a similar lexical choice and word order to those of the reference translations and, on the other hand, they penalize those hypotheses that show a different word order and vocabulary, even if they are valid translations. As a reaction to this, some researchers suggested taking advantage of lightweight linguistic knowledge such as stemming, use of synonyms and paraphrasing (Russo-Lassner et al. 2005, Zhou et al. 2006; Owczarzak et al. 2006).

#### 2.2.2.2 Lightweight Linguistic Knowledge Metrics

As a response to the weak points of those metrics not based on linguistic knowledge, some researchers have been using low-level linguistic knowledge, mainly morphological information, stemming, use of synonyms and paraphrasing. Next some of the most important metrics that use lightweight linguistic knowledge are presented.

- METEOR (Banerjee and Lavie 2005). METEOR is an F-measure based on unigram alignment. The original version of METEOR matched words in the

candidate translation with words in the reference translation by means of the following word-mapping modules in the following order: exact module, which maps two words if they are exactly the same; stem module, that maps two words if they show the same stem after applying the Porter stemmer (Porter 2001); synonym module, that matches two words if they show a relation of synonymy according to WordNet (Miller and Fellbaum 2007). If more than one alignment is found, METEOR selects the alignment for which the word order in the two strings is most similar. In addition, METEOR also applies a penalty which accounts for word order. However, authors did not stop here and continued improving this metric with more linguistic features, such as METEOR-NEXT (Denkowski and Lavie 2010) which allows word and phrase matches between the hypothesis and reference strings. Besides METEOR-NEXT includes a new mapping module which takes into account paraphrasing and a fragmentation penalty to account for gaps and differences in word order (Lavie and Agarwal 2007). Paraphrasing is done by using paraphrase tables and matching those phrases listed as paraphrases in these paraphrase tables. In order to build these paraphrase tables they used released bilingual corpora. The original version of METEOR contained three parameters: one for controlling the relative weight of precision and recall when computing the Fmean score; one controlling the shape of the penalty as a function of fragmentation; and finally, one for the relative weight assigned to the fragmentation penalty. In METEOR-NEXT these three parameters were tuned for fluency and adequacy, separately, and also for their combination. They were also tuned depending on the language evaluated: English, Spanish, French and German. Later METEOR-NEXT turned into METEOR 1.3 (Denkowsi and Lavie 2011), a version that includes a text normalizer (dealing with punctuation marks), filtered paraphrase tables (improving their precision), and function words list (distinguishing between content words, and therefore important words in the translation, and function words). The latest version of METEOR, METEOR Universal (Denkowski and Lavie 2014), covers previously unsupported languages by using automatically learned linguistic resources, mainly a function words list and

paraphrases, and combining them with a universal parameter for all languages, in opposition to the language-specific parameters used in previous versions.

- *Stochastic Iterative Alignment* (SIA) (Liu and Gildea 2006). This metric is based on a loose sequence alignment but improved with alignment scores, that is to say, it computes the string alignment score based on the gaps in the common sequence; stochastic word matching (it uses a soft matching based on the similarity between two words instead of using a stemmer or WordNet-related information); and an iterative alignment scheme, in other words, the string alignment continues until there are no more co-occuring words found between the hypothesis and reference translation.

- M-TER (Agarwal and Lavie 2008). This metric is an extended version of the edit-distance measure TER which uses the stemming and WordNet-based word mapping from METEOR.

- M-BLEU (Agarwal and Lavie 2008). Similar to the extension of the edit-distance measure TER, the authors also extended the well-known metric BLEU by means of using stemming and WordNet-based mapping modules from METEOR.

- *Assessment of Text Essential Characteristics* (ATEC) (Wong and Kit 2008/2010). This measure relies on two fundamental linguistic features: word choice and word order. As for word choice and word matching, ATEC is equipped with a module dealing with stemming and another one dealing with synonyms, by means of a WordNet-based (Wu and Palmer 1994) and a corpus-based measure (Landauer et al. 1998). Besides, the informativeness of the word is also taken into account in two ways: a) word-matches are weighed differently according to the informativeness of the word; b) unmatched words with a high degree of informativeness are also taken into account when quantifying the information missed. Regarding word order, ATEC uses three explicit features for word order: a) position distance, the closer the positions of a matched word in the candidate and reference are, the

better the match it is; b) order distance, which concerns the information flow of a sentence in the form of a sequence of matches; and c) the size of the phrase, longer phrases are given more credit to reward its valid word sequence.

- *Translation Edit Rate Plus* (TER-Plus/TERp) (Snover et al. 2009). TERp is an extension of TER that overcomes TER limitations by means of the addition of three new types of edit operations: stem matches, synonym matches and phrase substitutions. Besides, it also allows for a different cost depending on the type of match.

- INVWER and CDER (Leusch and Ney 2009). Both measures are described as edit-distance measures. *Inversion Word Error Rate* (INVWER) is a variant of WER which allows for block movements, although it does not demand complete and disjoint coverage of the source sentence. *Cover/Disjoint Error Rate* (CDER) (Leusch et al. 2006/2009) is a recall-oriented measure that models block reordering as an edit operation, with the restriction that block operations have to be bracketed. The authors described this measure using a Bracketing Transduction Grammar and sketched a polynomial-time algorithm for its calculation. Both measure use information on spelling and prefixes. On the other hand, these authors also proposed CD6P4ER, a linear combination of PER and CDER (see 2.2.2.4 for more details).

- *Modified BLEU, Enhanced Ranking metric* (AMBER) (Chen and Kuhn 2011; Chen et al. 2012). This measure is based on BLEU but incorporates recall, extra penalties and some text processing variants. The metric relies mainly on surface comparisons, although it also uses some linguistic knowledge, namely morphological knowledge on suffixes, roots and prefixes in the preprocessing stage. Besides, researchers used *tf-idf* (frequency-inverse document frequency)[7] to weigh n-grams differently according to their informative value.

[7] A numerical measure to reflect how important a word is to a document in a corpus.

- *Stanford Probabilistic Edit Distance Evaluation* (SPEDE) (Wang and Manning 2012). This metric makes prediction of translation quality by computing weighted edit distance. The authors used a probabilistic finite state machine, where state transitions corresponded to edit operations, in order to model weighted edit distance. The weights of the edit operations were then automatically learned in a regression framework. Besides, the use of a probabilistic Pushdown Automaton improves traditional edit-distance models as it allows for phrase shift and word swapping. Finally, this method also uses WordNet Synonyms and paraphrasing by means of the above TERp paraphrase table.

- Popović (2012) proposed BLOCKERRCATS (BΣER), ENXERRCATS (ENXΣER), WORD-BLOCKERRCATS (WBΣER) and XENERRCATS (XENΣER). This family of metrics is based on the classification and combination of translation errors. Translation errors are classified into five basic class error rates: a) INFER, translated words that have a correct base but a problem in the inflection, normalized over the hypothesis length; b) RER, words translated and occupying an incorrect position in the sentence, normalized over the hypothesis length; c) MISER, number of words which are missed in the MT output (they have not been translated), normalized over the reference length; d) EXTER, number of extra-words in the candidate translation, normalized over the hypothesis length; e) LEXER: number of words mistranslated in the candidate translation normalized over the hypothesis length. These errors are calculated at both word level and block level (a group of consecutive words labeled with the same error category). These 5 basic error rates were also explored in combination (see 2.2.2.4 for further details)

- BEER - BEtter Evaluation as Ranking – (Stanojević and Sima'an 2014). This metric combines adequacy features and ordering features. As regards adequacy features, Precision, Recall and the F1 score are calculated on matched function words, matched content words, match words of any type and matching of character n-grams (for size up to 6). As regards ordering

features they represent reordering as a permutation and then measure the distance to the ideal monotone permutation. This metric is also tuned for human judgements.

Although this type of metrics have proved to be better than those that do not use linguistic knowledge, they only take into account lexical similarities. As NLP improves, several tools are available to the scientific community which may account for other similarities beyond the lexical ones. In fact, when a human evaluator rates an MT sentence it does not only focus on the lexical units but on the adequacy and/or fluency of the whole sentence, which implies other dimensions of language (e.g. morphosyntax, phrase and sentence structure, meaning, etc.). Therefore, it seems natural and necessary that MT metrics do also take into account information at phrase and sentence level, making good use of tools providing syntactic and semantic information.

#### 2.2.2.3 Heavyweight Linguistic Knowledge Metrics

As stated in the paragraph above, the metrics described in section 2.2.2.2 work at lexical level, however these metrics fail in accounting for phrase and sentence structure, thus missing important syntactic information that should also be taken into account when evaluating MT (see Chapter 4) and penalizing translations that show a legitimate syntactic variation. Thus, the research community has developed other more sophisticated metrics that use richer linguistic knowledge at phrase and sentence level, such as PoS, constituent analysis, dependency analysis, SR labeling and textual entailment. The drawback to this approach is that using rich linguistic knowledge implies that these metrics are language-dependent and NLP tools for each of the languages analysed must be available. In this section, we first focus on those metrics using syntactic information and then we deal with those using semantic information.

- **Syntax-based metrics:**

  This group of metrics cover those measures that use any kind of morphosyntactic and/or syntax-oriented features, such as PoS and constituent and dependency analyses.

- *SyntacticTree Matching* (STM) and *Head-Word Chain Matching* (HWCM) (Liu and Gildea 2005). The STM metric works at constituent level and is based on the fractions of the subtrees that appear both in the hypothesis and reference strings. For each MT output, the fractions of subtrees with different depths are calculated and their arithmetic means is computed. Regarding HWCM, which works at dependency level, this measure compares head-word chains from both the hypothesis and reference dependency trees. According to the authors both syntax-based metrics outperform BLEU in terms of fluency.

- *BLEU's Associate with Tectogrammatical Relations* (BLEUÂTRE) (Mehay and Brew 2007). This measure uses syntactic word-word dependencies based on head-dependent relationships from parses of reference translations. Syntactic dependencies of the reference translations are flattened and compiled in bags of dependent words that must appear at the right and left of the headword, which enforces a partial linear order of dependents with respect to their heads. Therefore, there is no need to parse all candidates, just check the linear dependencies appearing in the candidates. By means of this approach possible ill-formed automatic candidate translations are avoided. BLEUÂTRE does not use synonymy, paraphrasing or inflectional morphological information. Although BLEUÂTRE is a syntax-based metric, results reported by authors state that the metric correlates better with human accuracy judgements than with fluency judgements.

- Owczarzak et al. (2007a/b) followed Liu and Gildea (2005)'s approach on dependency-based MT evaluation, but enhanced its performance by means of several steps: a) unlike Liu and Gildea who used non-labelled head-word sequences, they used an LFG parser which provides labels with the type of grammatical relation that exists between the head and its modifier; b) the use of these labels allows for partial matches: for the predicate-only dependencies, partial matching is allowed (two triples are considered identical even if only the head or only the modifier match), however, for the non-predicate dependencies, matches have to be complete; c) they used a

number of n-best parses (10-best parses would be optimal according to the authors) which help to reduce the amount of noise produced by the parser; d) to allow for lexical variation between the hypothesis and reference strings they used WordNet synonyms. In experiments reported in (Owczarzak et al. 2007a/b) authors claimed that their method correlates better with fluency human judgements than accuracy ones.

- *Surface Span Extension to Syntactic Dependency Precision-based MT Evaluatio*n (SEPIA) (Habash and Elkholy 2008). This measure uses dependency representation but it also includes surface span when calculating the evaluation score. The dependency span is the surface distance between two words that show a direct relation in a dependency tree. Thus, long-distance dependencies should receive a higher weight than short-distance dependencies, since according to the authors:

  > "we suspect that long-distance matches indicates a higher degree of grammaticality" (Habash and Elkholy 2008).

  Therefore, SEPIA seeks for capturing the grammaticality/fluency of a sentence.

- Giménez (2008a) proposed the SP (Shallow Parsing), CP (Constituent Parsing) and DP (Dependency Parsing) metrics. A family of metrics based on overlapping of certain linguistic features such as PoS, constituents and dependency relations, between the hypothesis and reference string. The SP metric can calculate overlapping over a particular type of PoS or chunk type, or over all PoS and chunk types. The CP metric calculates overlapping of PoS and phrase constituents (allowing for phrase embeddings, in opposition to SP metric) over a specific type or over all PoS and phrase consitutent types. Finally, the DP metric computes overlapping between words hanging at the same level, words directly hanging from terminal nodes and words ruled by non-terminal nodes. Similarly to SP and CP, it allows for overlap over a specific type of level, category or relation or all level, category and relation types. Sentences are annotated using the SVM tool (Giménez and

Márquez 2004), Freeling (Carreras et al. 2004) and Phreco (Carreras et al. 2005) in the case of PoS and chunk types, the MINIPAR parser (Linn 1998) for dependency parsing and Charniak-Johnson's Max-Ent reranking parser (Charniak and Jonson 2005) for constituent parsing.

- *Expected Dependency Pair Match* (EDPM) (Kahn et al. 2010). This method follows the approach started by Owczarzak et al. (2007a/b), but instead of using an LFG parser, a publicly available PCFG parser is used. Each sentence is analysed and transformed into a labeled syntactic dependency tree and then relations from each tree are extracted in the form of <Dependent, arc-Label, Head> subtree tuples, and compared. Several syntactic decompositions are explored resulting in 4-best type of decompositions which are combined in the final metric: 1) *dlh*, Dependent, arc-Label[8] and Head; 2) *lh*, arc-Label and Head; 3) *1g*, a simple measure of unigram precision and recall; 4) *2g*, a simple measure of bigram precision and recall.

- *DCU Dependency-based metric* (He et al. 2010). This measure is an extended version of Owczarzak et al. (2007a/b) in a different way. This means that the metric is based on dependency similarity but several improvements have been performed. First, the metric uses an adapted version of the Malt parser (Nivre et al. 2006) and instead of using n-best parses, only the 1-best parse is used. Second, stemming, synonymy and paraphrase information is added in order to allow for lexical variety. Third, the type of matches has been changed taking into account complete matches (same label, head and modifier), partial matches (same label and head but different modifier) and soft matches (same head and modifier but different label). Fourth, labels and matches are weighted in order to achieve a higher correlation with human judgements. Finally, a chunk penalty is introduced, following METEOR's string-based approach, in order to consider word order and fluency.

[8] Dependent refers to the modifier in the dependency relation, Arc-Label to the dependency relation label and Head to the head of the dependency relation.

- *Translation Evaluation of Sentences with Linear-programming-based Analysis* (TESLA) (Liu et al. 2010; Dahlmeier et al. 2011). A family of metrics (TESLA-M, TESLA-B and TESLA-F), which uses linguistic knowledge in the three versions. TESLA-M calculates the arithmetic average of F-measures between bags of n-grams – a multiset of weighted n-grams. So as to decide the weight that each n-gram receives, PoS are used to identify function or content words. Besides, information on lemmas is used, as well as WordNet which accounts for synonymy relations. TESLA-B uses the same information as TESLA-M, but it includes bilingual phrase tables to model phrase synonyms. TESLA-F, the most sophisticated of the metrics, additionally uses language models and a ranking support vector machine instead of simple averaging (as in TESLA-M and TESLA-B). This metric seeks good correlation with human judgements on ranking.

- Popović (2011) proposes POSF, MPF, WMPF. These metrics are aimed at using morphemes and PoS tags for n-gram based evaluation metrics. POSF, the PoS n-gram-based F-measure, is calculated on single units. It takes into account all PoS n-grams which appear both in the corresponding reference and the hypothesis. MPF calculates the Fscore on pairs of morphemes and PoS tags. Finally, WMPF calculates the F-score on word, morpheme and PoS n-grams. These measures require a PoS tagger of the target language, as well as a tool to split words into morphemes. These measures showed a good correlation with human judgements on adequacy, fluency and ranking.

- *Translation Error Categorization-based MT Quality Metric* (TerrorCat) (Fishel et al. 2012). This measure aims at quantifying translation quality based on the frequencies of different error categories. First, automatic error analysis is applied to the system outputs, providing the frequency of every error category for each sentence. Then, the Berkeley aligner (Liang et al. 2006) is applied to the whole set of reference-hypothesis pairs. Later, pairwise comparison of sentence pairs is achieved by means of an SVM classifier. There is also a model that allows for testing the number of errors according to the PoS tag, therefore using a lemmatizer and a PoS tagger.

- DepRef (Wu et al. 2013). This metric relies only on the REFerence DEPendency tree, which contains both lexical and syntactic information. The hypothesis segment remains unparsed in order to avoid parser errors propagation. In order to calculate the Fscore, the string of the hypothesis and dependency-based n-grams extracted from the reference dependency tree are used. There are two versions of this metric, one does not use external resources, whereas the other one uses stemming and synonymy relations. Most recent versions of this metric are the RED family (Wu et al. 2014), which follows a parametric approach and more information at lexical level (i.e. stemming, synonymy, function words and paraphrasing). In addition, different strategies in the parameters tuning are followed to obtain the different versions of the metric. DepRef showed good correlations with human judgements on ranking in the WMT13 Metrics Task.

The metrics reported in this section aim, mainly, at evaluating the grammaticality of the MT output, thus they are rather fluency-oriented than adequacy-oriented. However, a "grammatically" correct sentence does not imply that it is also "semantically" correct. In other words, a high score in fluency does not imply a high score in adequacy. And the other way round, a syntactically ill-formed sentence can still convey the meaning of the source sentence, as will be shown later in Chapter 6. As a reaction to this, the MT evaluation community has also become interested in those metrics which rely more on semantics.

- **Semantic-based metrics:**

The semantic level is probably the level which less metrics have been developed for. This is probably due to the difficulty that semantics poses to NLP and that although resources such as WordNet are widely used, they are just focused on lexical semantics. Dealing with sentence semantics is still a challenge for NLP, and just some automatic resources are available at this label (e.g. SR labeling, NEs recognition and identification, textual entailment, etc.). Metrics working at this level also use information related to discourse representation. Some of the most prominent metrics working at semantic level are described below.

- ‘NEE’ metric (Reeder et al. 2001). This measure was devoted to measuring MT quality over named entities.

- ‘NE’ metrics (Giménez 2008a). These metrics intend to capture similarities over named entities in the hypothesis and reference strings. They used named entities overlapping[9] and named entities matching[10] and they can calculate overlapping and matching over a specific type of NE or over all NE types.

- ‘SR’ metrics (Giménez 2008a). This family of metrics captures similarities between Semantic Roles in the hypothesis and those in the reference translation. These measures compute lexical overlapping – SR-Or(*) – and lexical matching – SR-Mr(*) – between Semantic Roles, taking into account Semantic Roles of a specific type or all Semantic Role types. Besides, another measure (SR-Or) reflects Semantic Role overlapping regardless of their lexical realization. In order to get Semantic Roles annotation for both hypothesis and reference strings, the SwiRL package (Surdeanu and Turmo 2005) is used.

- ‘DR’ metrics (Giménez 2008a). This is a family of metrics based on the Discourse Representation Theory (DRT) by Kamp (1981). This theory uses Discourse Representation Structures (DRS), which are a variation of first order predicate calculus. The hypothesis and reference translation are first analysed using C&C tools (Clark and Curran 2004), and once the strings are annotated with DRS, the DR metrics compare DRSs in the hypothesis and reference translation. In order to calculate the metrics score, lexical overlapping, morphosyntactic overlapping and the fraction of matching subpaths of a given length are computed. Similar to SR metrics, calculations are made over a specific type of DRS or over all DRS types.

- Stanford RTE (Recognition of Textual Entailment) system for MT evaluation (Padó et al. 2009). This measure approaches MT from the point

[9] Overlapping “*provides the proportion of items inside elements of the same type that have been successfully translated*” (Giménez 2008a).

[10] Matching is similar to overlapping but taking into account the relative order of the items.

of view of Textual Entailment, in other words, a correct MT output should be semantically equivalent to the reference translations; thus, both segments should entail each other. This measure uses two models: the first one is "a regularized linear regression model over entailment-motivated features that predict an absolute score for each reference hypothesis pair" (Padó et al. 2009); the second model considered is a regularized logistic regression model that predicts a weighted binary preference for each hypothesis pair. Entailment features are provided by Stanford entailment recognition system (MacCartney et al. 2006), which conducts a robust dependency-based linguistic analysis on both candidate and reference translations, aligns dependency graphs of the two strings and, finally, computes roughly 75 features over both strings.

- SemPOS (Kos and Bojar 2009; Macháček and Bojar 2011). Kos and Bojar (2009) designed this measure using Tecto MT framework (Žabokrtský et al. 2008) to assign a semantic PoS (Sgall et al. 1986) and *t-lemmas* (deep-syntactic lemmas) instead of surface word forms. They calculated overlapping of semantic PoS and used lemma information instead of surface word forms. Later, (Macháček and Bojar 2011) improved the metric so as to reduce their computational cost.

- SAGAN-STS (Semantic Textual Similarity) (Castillo and Estrella 2012). Similar to Padó et al. (2009), this metric uses Textual Entailment Technology to check whether the MT output and the reference translation are equal from a semantic point of view. SAGAN-STS is based on a semantic textual similarity engine which uses eight WordNet-based similarity measures in order to obtain the maximum similarity between two concepts, and on SenSim, a sentence level semantic metric (Castillo and Cardenas 2010).

- The MEANT metric, proposed by Lo et al. (2012), which uses Semantic Role Labeling (SRL) to identify similarities between the automatic and the reference translations. Unlike the SR metric (Giménez 2008a), MEANT uses structured Semantic Role representations which enable the metric to capture

the structural relations in semantic frames. In addition, MEANT also weighs SRs differently according to their importance to the adequate preservation of meaning. MEANT works as follows: firstly, both candidate and reference translations are semantically analysed with SRs; secondly, the semantic frames obtained in the hypothesis and reference translations are aligned according to the lexical similarity of the predicates; thirdly, for each pair of aligned semantic frames, the similarity of the Semantic Role fillers is determined by means of lexical similarity scores; fourthly, the Semantic Role fillers are aligned taking into account their lexical similarity; finally, a weighted F-score is calculated over the matching role labels of the aligned predicates and role fillers. MEANT uses ASSERT (Pradhan et al. 2004) for SRL purposes. In recent years, new versions of MEANT have been released whose main changes imply a) unsupervised tuning of weights, one for each Semantic Role (Lo and Wu 2013) in the UMEANT metric and b) developing a new semantic-role based metric without reference translations, the Xmeant metric (Lo et al. 2014) (see section 2.2.1).

Although some of these metrics proved extremely effective to evaluate MT quality (e.g. the SR metric got the first position at system level in the WMT07 (Callison-Burch et al. 2007) and both UMEAN and MEANT got third and fourth positions, respectively, at system level in the WMT13 (Macháček and Bojar 2013)), it is undeniable that they are more focused on the meaning of the sentence than on its grammaticality; thus, focusing on a partial aspect of translation.

According to most recent lines of research, the use of a single metric does not seem the best way to evaluate MT output, as each metric focuses on partial aspects of quality and the result is that strongly biased evaluations are generated. Actually, some of the metrics reported so far have been combined in order to improve their performance (Leusch and Ney 2009; Popović 2012). The main consequences of using a single metric evaluation are the following. Firstly, the comparison between systems is unfair. It is widely accepted that the result of comparing an SMT system with a rule-based MT system using only a metric based on lexical similarity will end in a much better result for the statistically-based system than for the rule-based one.

Secondly, the adjustment of parameters in the development of an SMT system following the results of a metric working at a specific level can end up in a strongly biased system. Finally, a good translation is the one that contains good lexis, good phrase and sentence structure and that transmits the meaning of the source sentence. All in all, a more holistic and compositional approach seems to be the most appropriate strategy to evaluate MT output. In order to achieve a good combination of metrics, several strategies have been developed. These strategies and several proposed metric combinations are addressed in the following section.

#### 2.2.2.4 Combination of Metrics

When dealing with metric combinations, a couple of points should be taken into account. Firstly, how to combine several metric scores into a single one; in this regard, following Giménez (2008a), a couple of approaches can be distinguished: *parametric* and *non-parametric*. In the parametric approach, each metric's contribution is individually weighed through a specific parameter, whereas in the non-parametric approach the contribution of each individual metric to the global score does not rely on any parameter. Secondly, how to evaluate the quality of a metric combination; in this sense two criteria can be followed: human likeness, which is the ability for the metric to distinguish between human translations and automatic ones; and human acceptability, in other words, the correlation with human assessments. Thus, the several approaches that deal with the combination of metrics have been divided in this section as follows:

a. Parametric approaches based on human likeness

b. Parametric approaches based on human acceptability

c. Non-parametric approaches

a. Parametric approaches based on human likeness.

In this type of approach each metric's contribution into the final score is individually weighted by means of an associated parameter and the metric aims at distinguishing human translations from automatic translations. The following combination fall within this type:

- Corston-Oliver et al. (2001) described a system which aimed at evaluating the fluency of MT, by means of classifiers which emulate the human ability to distinguish MT from human translations. The authors used perplexity measures and linguistic features related to branching properties of the parse, function word density, constituent length and other features at lexical level (e.g. accounting for "out of vocabulary words"[11]).

- Kulesza and Schieber (2004) extended Corston-Oliver et al. (2001) approach to take into account other aspects of quality rather than fluency alone. Instead of using decision trees they trained Super Vector Machines (SVM) to combine features from several well-known metrics (BLEU, NIST, WER and PER).

- Gamon et al. (2005) proposed a similar approach to that of Kulesza and Schieber (2004), although it assessed MT quality and fluency at sentence level without reference translations. They combined standard language model perplexity scores with class probability scores from an SVM classifier trained to distinguish MT from human translations by using linguistic analysis features (i.e. PoS, constituents, semantic features such as definiteness). This approach did not outperform BLEU when correlating with human judgements, however, they obtained good results when identifying the worst-translated sentences in a classification task.

b. Parametric approaches based on human acceptability.

Similar to the previous category, the contribution of each metric to the final score is optimized by a specific parameter, although in this case the metric aims at satisfactorily correlating with human judgements on MT quality. Some of the metrics included in this category are described below:

- Akiba et al. (2001) introduced a method based on human acceptability as a multiclass classification task. The authors trained decision tree classifiers on a set of edit-distance features using a combination of lexical,

[11] Words that do not occur in the training data.

morphosyntactic and lexical semantic information (i.e. word-form, stem, PoS and semantic classification from a Thesaurus).

- Paul et al. (2007) proposed an approach which assessed several aspects of MT quality such as fluency, adequacy and acceptability. The authors trained SVM classifiers to combine the scores obtained from several lexical metrics (BLEU, NIST, METEOR, GTM, WER, PER and TER). However, none of the metrics used is based on syntax, therefore missing information at phrase and sentence level.

- Albrecht and Hwa (2007a/b) revised the SVM-classifier approach and proposed a regression-based approach to combine metrics with and without human references. They used four kinds of features to train their regression-based model: a) string-based metrics over references (those used in Kulescha and Schieber (2004)) with the addition of METEOR, ROUGE-inspired features and ROUGE-L); b) syntax-based metrics over references (HWCM and STM); c) string-based metrics over corpus; and d) syntax-based metrics over corpus.

- Ye et al. (2007) introduced a new approach based on ranking instead of classification. They used a ranking-SVM algorithm to rank candidate translations depending on several features: n-gram based features, dependency features (reduced to 5 dependency structure types) and translation perplexity according to a reference language model.

- Liu and Gildea (2007) proposed a new method which is based on maximum correlation training, in other words, the weight of contribution of each metric to the overall score is adjusted in order to maximize the level of correlation with human assessments at segment level. Besides, this combination of metrics also uses features from the source sentence: source sentence constrained n-gram precision and source-sentence reordering agreement.

- Leusch and Ney (2009) tried CD6P4ER, a linear combination of 0.4 PER and 0.6 CDER (see the description of PER in section 2.2.2.1 and CDER in

section 2.2.2.2) which, in their experiments, improved the correlation with human judgements on adequacy.

- Yang et al. (2011) followed SVM regression to assess MT quality on adequacy and fluency, and accuracy in pairwise comparison. They used linguistic features at word level (quality of content words and cognate words matching), at phrase level (by means of constituent parsing) and at sentence level (length comparison between the candidate translation and the source segment and parser score). Their experiments achieved a comparable or better correlation than those approaches based on rich linguistic features and reduced the risk of over-fitting.

- Specia and Giménez (2010) proposed a method that combines confidence estimation features and reference based metrics and a learning mechanism based on human annotations.

- Rios et al. (2011) described TINE, metric which aims at evaluating MT adequacy. TINE combines a lexical matching component and an adequacy component. The lexical matching component compares bags-of-words without any linguistic annotation (only word-forms and stems are used). The adequacy component uses ontologies to align predicates, Semantic Roles to align arguments, and finally it matches arguments using distributional semantics. Both components are weighed in order to correlate better with human judgements.

- Popović (2012) described 5 error rates (INFER, RER, MISER, EXTER and LEXER – see section 2.2.2.2 for further details) which were combined in: a) BLOCKERRCATS, sum of block level error rates; b) ENXERRCATS, linear interpolation of word level and block level class error rates optimized for translation from English; c) WORDBLOCKERRCATS, arithmetic mean of word and block level error rates; d) XENERRCATS, linear interpolation of word level and block level class error rates optimized for translation into English.

- Gautam and Bhattacharyya (2014) proposed Layered, a combination of metrics at lexical, syntactic and semantic layers. At lexical level BLEU is used; at syntactic level three metrics are considered (Hamming score (Hamming 1950), Kendall's Tau distance score (Kendall 1938/1955) and the Spearman rank score (Spearman 1904)) which take care of the reordering of words within the sentence. Finally, at semantic level they used two metrics Shallow semantic score and Deep semantic score, which are based on the concept of Textual Entailment. The former uses dependency relations provided by the Standford parser and the latter relies on the UNL[12] dependency generator. SVM-rank was used to learn the parameters/weights for each metric.

- DiscoTK metrics (Joty et al. 2014). This family of metrics is based on discourse representation. They compare discourse trees in the hypothesis and reference segments, computing similarities by means of convultion kernels. There are three main metrics: DiscoTK-light, which uses five different transformations and augmentations of discourse trees representations and combines their kernel scores into a single score; DiscoTK-party, which combines DiscoTK-light with other metrics available in the Asiya toolkit (Giménez and Márquez 2010a; González and Giménez 2014); finally, DiscoTK-party-tuned, which tunes the weights of the metrics using human judgements in a learning-to-rank framework.

c. Non-parametric approaches.

In opposition to the previously mentioned approaches, in the non-parametric approach the contribution of each individual metric to the global score does not rely on any parameter. Some of the metrics within this approach are described below.

- Chang and Ng (2008) proposed MAXSIM, a metric based on precision and recall, which allows for synonym matching and weighs the matches found. This measure combines linguistic features at two different levels: a) n-gram

[12] http://www.undl.org/unlsys/unl/unl2005/UW.htm

information (lemma and PoS match, lemma match and bipartite graph match using WordNet); b) dependency relations (dependency match of two functions – subject and object). In order to match n-grams they used the linguistic features mentioned above to match unigrams, bigrams and trigrams. The final score is obtained as follows: first, unigrams in the candidate translation are matched to the ones in the reference translation, precision and recall are calculated based on the matches and then combined into a single Fmean unigram score; second, a similar process is followed to calculate the Fmean of matched bigrams and trigrams; finally, the three Fmean scores (unigram, bigram and trigram scores) are simply averaged.

- Giménez and Márquez (2010a/b) proposed combining a set of individual metrics which worked at different levels (i.e. lexical, syntactic and semantic) by means of uniformly-averaged linear combination (ULC.) They used a large number of well-known metrics such as BLEU, NIST, GTM, METEOR (and its variants), ROUGE (and its variants), TER and others developed by the authors (see section 2.2.2.3, syntax-based and semantic-based metrics). Scores obtained from individual metrics are then averaged into a single measure of quality, without using any kind of parameters of machine learning techniques. The combination of metrics that they tested among different scenarios and that proved the most effective is the following one:

  $M=\{ROUGE_W, METEOR_{sy}, DP\text{-}HWCM_c, DP\text{-}HWCM_r, DP\text{-}O_r(*), CP\text{-}STM_4, SR\text{-}O_r(*), SR\text{-}M_r(*), SR\text{-}O_r, DR\text{-}O_r(*), DR\text{-}O_{rp}(*)\}$

  This combination of metrics proved really effective and outperformed all individual metrics in the WMT08 and WMT09 shared tasks. The drawback to this method is that it does not capture the importance of each linguistic feature, due to the diversity of metrics and the uniform weights used to aggregate linguistic features.

- González et al. (2014), following Giménez and Márquez (2010b), proposed two metrics: UPC-STOUT and UPC-IPA which use a large combination of metrics working at different levels combined by means of ULC. This wide

set of metrics involve metrics that disregard linguistic information, metrics using linguistic information at lexical level, syntactic level and semantic level. In addition, some of these metrics are source-based metrics, in other words, they compare the hypothesis segment with the source segment, disregarding reference translations.

Nowadays most of the measures participating in MT evaluation campaigns combine different metrics and information and they seem to obtain good results. It is undeniable that most of them follow a parametric approach and use machine learning techniques to reach the best performance. The drawback to this methodology is that they need a large amount of data to conduct their training.

### 2.3 Summing Up

In the last decade a wide range of MT evaluation metrics has been developed and great efforts are still made in order to improve already existing MT metrics and develop new ones; proof of this is the high number of participants in the last WMT14 shared task, where 23 metrics from 12 different research groups participated (Macháček and Bojar 2014).

From the MT metrics described in this chapter some of them do not use linguistic information at all, such as BLEU or NIST, and in fact, they are the most frequently used metrics because they are easy to use, fast and language independent. However, as stated in section 2.2.2.1, n-gram-based systems have a tendency to favour statistically-based systems over rule-based ones. As a reaction, metrics using linguistic information have been developed. Some of these metrics use lightweight linguistic information (e.g. synonyms, stemming, paraphrasing, difference between content and function words, etc.) such as METEOR, TERp, AMBER or SPEDE. They have proved quite effective – METEOR, for example, is widely used and shows good correlation with human judgements – however, they still disregard other linguistic similarities beyond lexical ones, such as sentence structure. As a consequence, taking advantage of improvements in NLP and the new resources and tools available, more sophisticated metrics, using heavyweight linguistic information (such as PoS tagging, constituent and dependency parsing or Semantic Role labeling) have been developed. Some of the most well-known

syntax-based metrics are Liu and Gildea (2005)'s SMT and HWCM metrics, Owczarzak et al. (2007a/b)'s approach, whereas in the case of SR-based metrics, Giménez (2008a)'s SR metrics and Lo et al. (2012)'s MEANT family are some of the most reknown. These types of metrics account for the syntactic and semantic sentence structure, respectively. Therefore, they are still focused on partial aspects of translation and the MT community, especially since ranking evaluations started to be used, have tried to reach a more holistic approach that can account for MT quality in general. In response to this need, researchers have been working, especially for the last five years, in the combination of metrics. As explained in this chapter, metrics can be combined following a parametric or a non-parametric approach. Most of the metric combinations developed fall under this parametric approach (e.g. Layered and DiscoTK metrics), and a wide range of them use machine learning techniques in order to optimize their measures so that they correlate well with human judgements on ranking. On the other hand, other researchers have decided to use a more simple approach, disregarding parameters and machine learning techniques, such as MASXIM and ULC.

After this last decade, it seems clear that linguistic information plays a crucial role in the evaluation of MT output and that combining linguistic information at different levels is the most successful approach to MT evaluation. This has recently been confirmed by Joti et al. (2014) and their DiscoTK family of metrics, a family of metrics that uses linguistic information at very different levels and which proved to correlate well with human judgements on ranking, since they occupied the first positions in the WMT14 shared task at both segment and system level evaluations. Nonetheless, some of the criticism that this type of metrics could receive is that they use such a large number of metrics that it is difficult to know their influence on the evaluation. In addition, these metrics are usually trained to correlate well with human ranking of sentences; however, this type of evaluation does not provide information on more partial aspects of translation which might be useful to developers in order to know which aspects of their systems must be improved and how they should do it. All in all, we consider that a more qualitative approach, going beyond correlations, could be of good help to developers aiming at improving their systems. An analysis of the source and target language considering their key features and focused on translation errors may lead to the development of a linguistically-enhanced MT metric working at different linguistic

levels, which does not require a large amount of data to obtain optimum results and which can be easily adapted to different evaluation tasks (i.e. adequacy, fluency or ranking).

# Chapter 3. Methodology

The present study pursues the aim of shedding some light on the suitability, influence and combination of linguistic information to evaluate MT output, especially by highlighting the effectiveness and benefits of a more qualitative approach based on linguistic analysis. In order to conduct our research, an empirical approach has been followed. In other words, we have mainly relied on data and experiments in order to formulate and confirm our hypotheses. Regarding the data analysed, the approach followed by our linguistic analysis was initially a corpus-based approach (Tognini-Bonelli 2001), because there were some linguistic phenomena that we had already expected to find in our data, such as the use of synonymy or typical mistakes related to noun-adjective agreement in Spanish. However, while the analysis was being carried out, some new phenomena arose, leading to a more corpus-driven approach. The linguistic analysis of both MT output and reference segments was conducted with the following aims: a) check the relevance of that linguistic information used so far by MT metrics from a linguistic point of view; b) try to find out whether other linguistic traits not currently used in any MT metrics could also help to improve automatic MT evaluation; c) try to determine those linguistic characteristics that are more relevant according to the type of evaluation (i.e. adequacy, fluency), considering both MT errors and positive linguistic phenomena that must be analysed as correct linguistic characteristics. Once the linguistic analysis was conducted, and taking into account the previous study of the MT metrics presented in Chapter 2, several hypotheses were formulated (section 1.1). In order to confirm them and explore the relevance of linguistic information, an MT evaluation metric (hereafter referred as "metric"), VERTa, was developed and experiments were carried out. Our metric is a linguistically-motivated metric since it is based on linguistic knowledge in opposition to those metrics that do not use linguistic information at all, such as WER (Nieβen et al. 2000), BLEU (Papineni et al. 2001) and NIST (Doddington 2002). In addition, our metric seeks the way of combining linguistic information at different levels in order to provide a more holistic evaluation, which is opposed to some of the most well-known and widely-spread metrics, which focus on a more partial evaluation such as the METEOR family (Denkowski and Lavie 2010), which works at lexical level; Liu and Hildea (2005)'s

proposal, which uses constituents information; the metrics proposed by Owczarzack et al. (2007a/b) and He et al. (2010), respectively, that use information based on dependency relations; and MEANT and UMEANT (Lo et al. 2012/2014), a method that uses Semantic Role labeling. VERTa follows a linguistic approach based on the linguistic analyses previously performed, in opposition to other metrics that also combine different types of information either trying to combine several metrics following a machine-learning approach (Leusch and Ney 2009; Albrecht and Hwa 2007a/b), or focusing on the combination of several metrics working at different levels (i.e. lexical, syntactic and semantic) in a more straightforward way and without much linguistic analysis (Chang and Ng 2008; Giménez and Márquez 2010a/b). The drawback to those approaches is that the former needs a large amount of data to obtain reliable results, whereas the latter does not capture the importance of each linguistic feature, due to the diversity of metrics and the uniform weights used to aggregate linguistic features. VERTa overcomes these two drawbacks because it does not require a large amount of data, since no machine-learning methods are used, and in addition, we use a smaller range of linguistic features which are easier to control and combine depending on the type of evaluation performed thanks to our initial linguistic analysis performed as well as the experiments and detailed post-analyses conducted. All this will be detailed in Chapter 5 devoted to the metric description, Chapter 6 and Chapter 7 aimed at Experiments on Adequacy and Fluency, respectively, and Chapter 8 on the Meta-evaluation of VERTa.

The current chapter covers the methodology used in our research. Firstly, the steps taken in our research are described (section 3.1); secondly, the resources and tools used in our study and to develop VERTa are detailed (section 3.2); finally, a summary of this chapter is provided (section 3.3).

### 3.1 Steps Followed to Conduct our Research

Our first hypothesis assumes that a linguistic analysis, indicated at the beginning of this chapter, can help to clarify what linguistic features should be used and how they should be combined to evaluate MT output. Therefore, a linguistic analysis of both MT output and reference segments was first carried out with the aim of highlighting not only those linguistic errors that an automatic MT evaluation metric must identify, but also those

positive linguistic features that must be taken into account, identified and treated as correct linguistic phenomena (see Chapter 4). So as to perform this linguistic analysis we used the English development corpus for adequacy (see section 3.2.1.1) and the Spanish corpus (see section 3.2.1.5). Once the linguistic analysis was conducted and in order to confirm our hypotheses and check whether those linguistic phenomena and traits identified in the analysis were helpful to evaluate MT output, we designed and implemented a linguistically-motivated MT metric, VERTa. So as to develop this metric, we first selected and evaluated those linguistic resources and tools that would be helpful to deal with the linguistic phenomena found. In order to deal with lexical semantics, we considered that WordNet 3.0 (Fellbaum 1998) was the most effective resource we could use since it provides a wide coverage for both English and Spanish and is used by most MT metrics. As for the tools used to syntactically annotate and parse the English corpus, and since there is a wide variety of NLP tools available for this language, both a quantitative and qualitative evaluation of the most well-known constituency and dependency parsers was performed (Comelles et al. 2010). After this evaluation, we opted for the Standford CoreNLP suite (Manning et al. 2014) to parse our English data since it proved the most effective parser to analyse dependency relations and we found it very convenient that we could get information about different types of analysis and annotation using the very same tool. As for the tools used to annotate and parse our Spanish corpus, we selected Freeling (Padró and Stanilovsky 2012) because it integrates different types of analyzers and being a knowledge-based tool which does not require any training, it seemed the most useful tool to analyse our rather small Spanish corpus.

Once the linguistic analysis was performed and the NLP tools were selected, we designed the architecture of our metric (see Chapter 5) and developed a first version of VERTa, consisting mainly of the Lexical, Morphological, Dependency, N-gram, Semantic and Language Model (LM) Modules for English[13]. Next, experiments to evaluate adequacy in English were performed (Chapter 6) using the English development corpus (see section 3.2.1.1). Our second and third hypotheses state that

[13] The Semantic and Language Model Modules are not available for Spanish.

a) Addressing fluency and adequacy evaluations separately would help to easily identify the use and suitability of linguistic features. Therefore, linguistic features would be more or less appropriate depending on the type of evaluation, either adequacy or fluency.

b) Studying different evaluation tasks might not be only useful to identify which linguistic features are more or less appropriate depending on the type of evaluation (adequacy or fluency) but also how they should be combined.

In order to investigate these hypotheses, firstly, the linguistic features included in each module were tested separately in order to check their suitability and their importance to evaluate adequacy; secondly, the best way to combine those linguistic features inside each module was examined; thirdly, the performance of the module itself to evaluate adequacy was also tested. Finally, the best combination of modules was checked by means of a system of weights. Modules' weights were first assigned manually, following linguistic criteria; although later in order to calculate an upper-bound for the weight tuning, all possible weight combinations were tuned automatically using a 0.01 step. On a second stage, experiments to evaluate fluency in English (see Chapter 7) were conducted using the English development corpus (see section 3.2.1.2). We followed the same process and steps as in the experiments performed to evaluate adequacy.

These experiments were always carried out at segment level, since we were interested in a fine-grained analysis and the evaluation of both adequacy and fluency was based on scores instead of ranking, since we consider scores to be more informative for our research. In addition, Pearson Correlation Coefficient (1914/1924/1930) was used to correlate scores provided by the metric with those provided by human judges. Traditionally, researchers use the correlation of the metrics' score with human judgements as a way to measure the performance of their metrics and to check the suitability of the features used. In our case, we also used the information provided by correlating VERTa's scores with human judgements as a clue to know whether we were making progress and we were advancing in the correct way. However, since our aim was not developing an MT metric but finding and checking the suitability of linguistic information in order to evaluate MT, besides using information provided by correlations

as a guide, we also performed a qualitative and detailed analysis of the metric's output every time linguistic features were added and/or combined. This analysis was possible due to the fact that VERTa does not only provide a score per segment but it also provides an XML file where linguistic features used in each module and their corresponding matches can be traced, as shown in the figures below.

The information provided in these files specifies the hypothesis and reference segments compared (see Figure 1), as well as all the relevant linguistic information that will be used to compare them.

**Hypotesis Sentence**

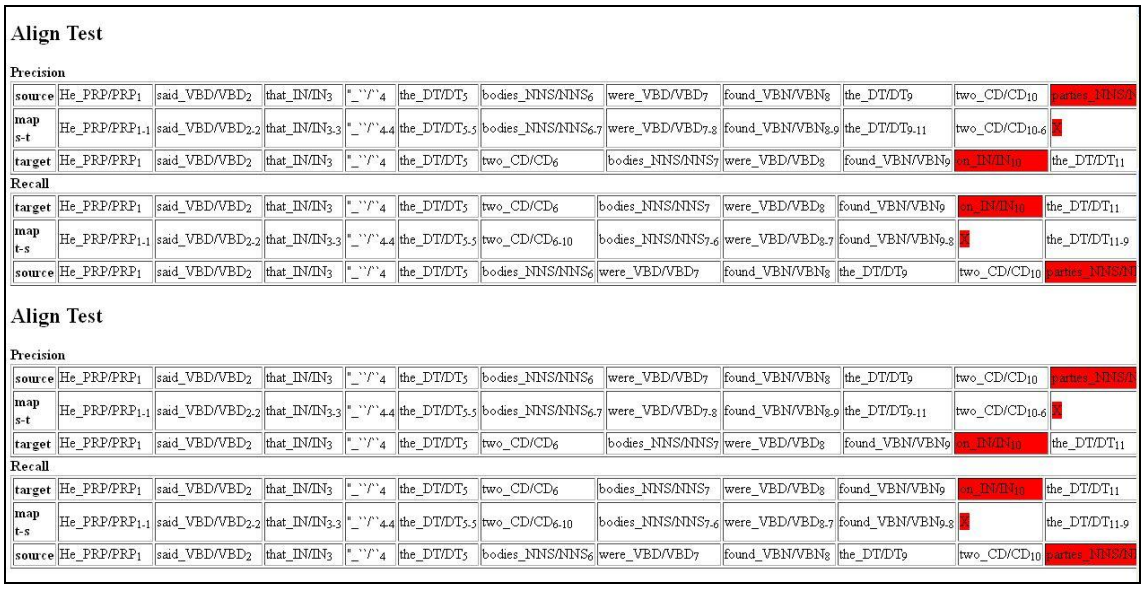
| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| WORD | He | WORD | said | WORD | that | WORD | " | WORD | the | WORD | bodies | WORD | were | WORD | found | | | | |
| CONL | 0 | CONL | 0 | CONL | 0 | CONL | 0 | CONL | 0 | CONL | 0 | CONL | 0 | CONL | 0 | | | | |
| WNSS | 0 | WNSS | 0 | WNSS | 0 | WNSS | 0 | WNSS | 0 | WNSS | B-E:PER_DESC | WNSS | 0 | WNSS | 0 | | | | |
| DEPLABEL | nsubj | DEPLABEL | _ | DEPLABEL | complm | DEPLABEL | punct | DEPLABEL | det | DEPLABEL | nsubjpass | DEPLABEL | auxpass | DEPLABEL | ccomp | | | | |
| WSJ | 0 | WSJ | B-verb.communication | WSJ | 0 | WSJ | 0 | WSJ | 0 | WSJ | B-noun.body | WSJ | 0 | WSJ | B-verb.percept | | | | |
| DEPHEAD | 2 | DEPHEAD | 0 | DEPHEAD | 8 | DEPHEAD | 8 | DEPHEAD | 6 | DEPHEAD | 8 | DEPHEAD | 8 | DEPHEAD | 2 | | | | |
| ID | 1 | ID | 2 | ID | 3 | ID | 4 | ID | 5 | ID | 6 | ID | 7 | ID | 8 | | | | |
| POS | PRP | POS | VBD | POS | IN | POS | `` | POS | DT | POS | NNS | POS | VBD | POS | VBN | | | | |
| LEMMA | he | LEMMA | say | LEMMA | that | LEMMA | " | LEMMA | the | LEMMA | body | LEMMA | be | LEMMA | found | | | | |
| SPOS | PRP | SPOS | VBD | SPOS | IN | SPOS | `` | SPOS | DT | SPOS | NNS | SPOS | VBD | SPOS | VBN | | | | |

**Reference Sentence**

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| WORD | He | WORD | said | WORD | that | WORD | " | WORD | the | WORD | two | WORD | bodies | WORD | were |
| CONL | 0 | CONL | 0 | CONL | 0 | CONL | 0 | CONL | 0 | CONL | 0 | CONL | 0 | CONL | 0 |
| WNSS | 0 | WNSS | 0 | WNSS | 0 | WNSS | 0 | WNSS | 0 | WNSS | B-N:CARDINAL | WNSS | B-E:PER_DESC | WNSS | 0 |
| DEPLABEL | nsubj | DEPLABEL | ccomp | DEPLABEL | complm | DEPLABEL | punct | DEPLABEL | det | DEPLABEL | num | DEPLABEL | nsubjpass | DEPLABEL | auxp |
| WSJ | 0 | WSJ | B-verb.communication | WSJ | 0 | WSJ | 0 | WSJ | 0 | WSJ | B-adj.all | WSJ | B-noun.body | WSJ | 0 |
| DEPHEAD | 2 | DEPHEAD | 25 | DEPHEAD | 9 | DEPHEAD | 9 | DEPHEAD | 7 | DEPHEAD | 7 | DEPHEAD | 9 | DEPHEAD | 9 |
| ID | 1 | ID | 2 | ID | 3 | ID | 4 | ID | 5 | ID | 6 | ID | 7 | ID | 8 |
| POS | PRP | POS | VBD | POS | IN | POS | `` | POS | DT | POS | CD | POS | NNS | POS | VBD |
| LEMMA | he | LEMMA | say | LEMMA | that | LEMMA | " | LEMMA | the | LEMMA | two | LEMMA | body | LEMMA | be |
| SPOS | PRP | SPOS | VBD | SPOS | IN | SPOS | `` | SPOS | DT | SPOS | CD | SPOS | NNS | SPOS | VBD |

**Figure 1** XML trace of the beginning of a hypothesis and a reference segments

They also show the information used in each module, starting by the Lexical and Morphological Modules (see Figure 2) and specifying the matches in these two modules and highlighting those elements that do not match in red.



Align Test

Precision

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| source | He_PRP/PRP$_1$ | said_VBD/VBD$_2$ | that_IN/IN$_3$ | "_``/``$_4$ | the_DT/DT$_5$ | bodies_NNS/NNS$_6$ | were_VBD/VBD$_7$ | found_VBN/VBN$_8$ | the_DT/DT$_9$ | two_CD/CD$_{10}$ | parties_NNS/N |
| map s-t | He_PRP/PRP$_{1-1}$ | said_VBD/VBD$_{2-2}$ | that_IN/IN$_{3-3}$ | "_``/``$_{4-4}$ | the_DT/DT$_{5-5}$ | bodies_NNS/NNS$_{6-7}$ | were_VBD/VBD$_{7-8}$ | found_VBN/VBN$_{8-9}$ | the_DT/DT$_{9-11}$ | two_CD/CD$_{10-6}$ | |
| target | He_PRP/PRP$_1$ | said_VBD/VBD$_2$ | that_IN/IN$_3$ | "_``/``$_4$ | the_DT/DT$_5$ | two_CD/CD$_6$ | bodies_NNS/NNS$_7$ | were_VBD/VBD$_8$ | found_VBN/VBN$_9$ | on_IN/IN$_{10}$ | the_DT/DT$_{11}$ |

Recall

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| target | He_PRP/PRP$_1$ | said_VBD/VBD$_2$ | that_IN/IN$_3$ | "_``/``$_4$ | the_DT/DT$_5$ | two_CD/CD$_6$ | bodies_NNS/NNS$_7$ | were_VBD/VBD$_8$ | found_VBN/VBN$_9$ | on_IN/IN$_{10}$ | the_DT/DT$_{11}$ |
| map t-s | He_PRP/PRP$_{1-1}$ | said_VBD/VBD$_{2-2}$ | that_IN/IN$_{3-3}$ | "_``/``$_{4-4}$ | the_DT/DT$_{5-5}$ | two_CD/CD$_{6-10}$ | bodies_NNS/NNS$_{7-6}$ | were_VBD/VBD$_{8-7}$ | found_VBN/VBN$_{9-8}$ | | the_DT/DT$_{11-9}$ |
| source | He_PRP/PRP$_1$ | said_VBD/VBD$_2$ | that_IN/IN$_3$ | "_``/``$_4$ | the_DT/DT$_5$ | bodies_NNS/NNS$_6$ | were_VBD/VBD$_7$ | found_VBN/VBN$_8$ | the_DT/DT$_9$ | two_CD/CD$_{10}$ | parties_NNS/N |

Align Test

Precision

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| source | He_PRP/PRP$_1$ | said_VBD/VBD$_2$ | that_IN/IN$_3$ | "_``/``$_4$ | the_DT/DT$_5$ | bodies_NNS/NNS$_6$ | were_VBD/VBD$_7$ | found_VBN/VBN$_8$ | the_DT/DT$_9$ | two_CD/CD$_{10}$ | parties_NNS/N |
| map s-t | He_PRP/PRP$_{1-1}$ | said_VBD/VBD$_{2-2}$ | that_IN/IN$_{3-3}$ | "_``/``$_{4-4}$ | the_DT/DT$_{5-5}$ | bodies_NNS/NNS$_{6-7}$ | were_VBD/VBD$_{7-8}$ | found_VBN/VBN$_{8-9}$ | the_DT/DT$_{9-11}$ | two_CD/CD$_{10-6}$ | |
| target | He_PRP/PRP$_1$ | said_VBD/VBD$_2$ | that_IN/IN$_3$ | "_``/``$_4$ | the_DT/DT$_5$ | two_CD/CD$_6$ | bodies_NNS/NNS$_7$ | were_VBD/VBD$_8$ | found_VBN/VBN$_9$ | on_IN/IN$_{10}$ | the_DT/DT$_{11}$ |

Recall

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| target | He_PRP/PRP$_1$ | said_VBD/VBD$_2$ | that_IN/IN$_3$ | "_``/``$_4$ | the_DT/DT$_5$ | two_CD/CD$_6$ | bodies_NNS/NNS$_7$ | were_VBD/VBD$_8$ | found_VBN/VBN$_9$ | on_IN/IN$_{10}$ | the_DT/DT$_{11}$ |
| map t-s | He_PRP/PRP$_{1-1}$ | said_VBD/VBD$_{2-2}$ | that_IN/IN$_{3-3}$ | "_``/``$_{4-4}$ | the_DT/DT$_{5-5}$ | two_CD/CD$_{6-10}$ | bodies_NNS/NNS$_{7-6}$ | were_VBD/VBD$_{8-7}$ | found_VBN/VBN$_{9-8}$ | | the_DT/DT$_{11-9}$ |
| source | He_PRP/PRP$_1$ | said_VBD/VBD$_2$ | that_IN/IN$_3$ | "_``/``$_4$ | the_DT/DT$_5$ | bodies_NNS/NNS$_6$ | were_VBD/VBD$_7$ | found_VBN/VBN$_8$ | the_DT/DT$_9$ | two_CD/CD$_{10}$ | parties_NNS/N |

**Figure 2** XML trace for the matches in the Lexical and Morphological Modules

The dependency triples matches established by the Dependency Module can also be traced (see Figure 3). In columns "source" and "target" the triples for the hypothesis and reference segments are shown and in the column "pattern" the types of match between the hypothesis and reference segments are established, being (X,X,X) an exact match, (X,O,X) a No_head match, (O,X,X), No_label match and (X,X,O) No_mod match. In addition, next to each type of match their corresponding weight is stated.

## Dependency triples

### Precision

| # | Max score | Score | source | target | Pattern |
|---|---|---|---|---|---|
| 1 | 1.0 | 1.0 | nsubj(said:2,He:1) | nsubj(said:2,He:1) | ( [nsubj] auto_top setw:1.0,X,X) : 1.0 |
| 2 | 0.0 | -1.0 | _(TOP:0,said:2) | NO MATCH | nomatch |
| 3 | 1.0 | 1.0 | complm(found:8,that:3) | complm(found:9,that:3) | (X,X,X) : 1.0 |
| 4 | 1.0 | 1.0 | punct(found:8,":4) | punct(found:9,":4) | (X,X,X) : 1.0 |
| 5 | 1.0 | 1.0 | det(bodies:6,the:5) | det(bodies:7,the:5) | (X,X,X) : 1.0 |
| 6 | 1.0 | 1.0 | nsubjpass(found:8,bodies:6) | nsubjpass(found:9,bodies:7) | (X,X,X) : 1.0 |
| 7 | 1.0 | 1.0 | auxpass(found:8,were:7) | auxpass(found:9,were:8) | (X,X,X) : 1.0 |
| 8 | 1.0 | 1.0 | ccomp(said:2,found:8) | ccomp(said:2,found:9) | (X,X,X) : 1.0 |
| 9 | 1.0 | 0.9 | det(area:13,the:9) | det(outskirts:12,the:11) | (X,O,X) : 0.9 |

**Figure 3** XML trace for the Dependency Module

The N-gram Module and its matches are also covered (see Figure 4). Those words that belong to an n-gram match appear in green, whereas those that have been disregarded are not coloured. Moreover, in these XML files information regarding the match length and the number of matched n-grams is also provided.

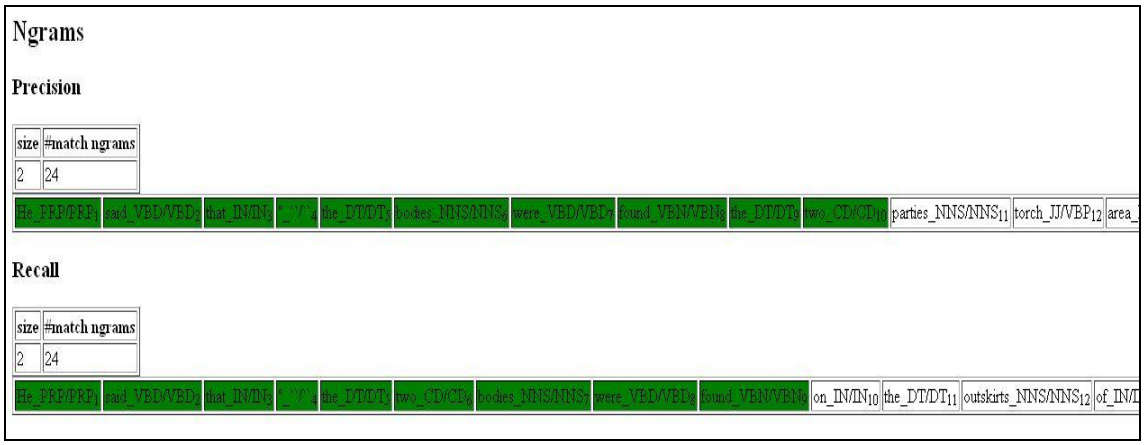



**Figure 4** XML trace corresponding to the N-gram Module

Finally, Figure 5 shows the XML trace corresponding to the features in the Semantic Module (i.e. NEs Recognition, NEs Linking, Sentiment analysis and Time Expressions). The semantic features in Figure 5 correspond to the hypothesis and reference segments below:

HYP: *In the 1950's and 1960's of the 20th century, targeting several church frequented by the state of Alabama in the soda.*
REF: *In the 1950's and 1960's, many churches frequented by blacks were targeted in the state of Alabama.*

In Figure 5, Named Entities corresponds to NEs recognition, Linked Named Entities corresponds to those NEs that can be linked through Wikipedia, Sentiment corresponds to the Sentiment analysis – whether the sentence is positive or negative –, and finally, Timex corresponds to the Time Expressions.

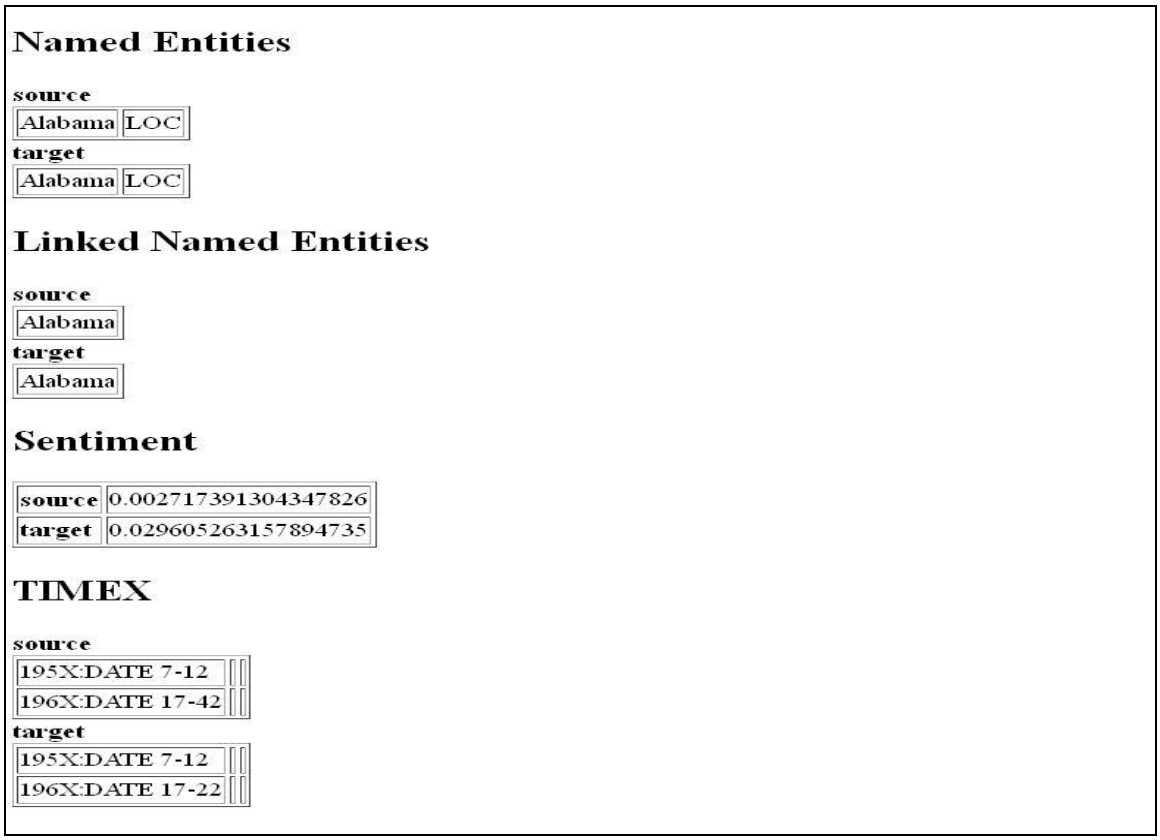



**Figure 5** XML trace corresponding to the Semantic Module

Therefore, every time we added and/or combined a new linguistic feature, we first checked the correlation with human judgements as a hint to see whether they improved or worsened. In both cases 40 segments were selected: 20 of them had achieved better scores whereas the other 20 had obtained worse scores in relation to human judgements. These segments were analysed in depth so that we could study how linguistic features influenced our metric for better or for worse, and we could make a final decision on the use of such features.

After performing these experiments and checking the suitability of the linguistic information used and how it had to be used and combined, VERTa's parameters were adjusted and a new and updated version of the metric was ready to be used. With this updated version and for the sake of comparison, a meta-evaluation of the metric for adequacy, fluency and MT quality was conducted, as well as a comparison to some of the best-known and widely-used MT metrics (Chapter 8), showing that it outperformed them all when adequacy and fluency were assessed.

Finally, our last hypothesis addresses the fact that porting a linguistically-enhanced MT metric to a new language may involve studying the main key features of that language and reflecting them on how linguistic features are used in the metric. To confirm this last hypothesis, one more experiment was conducted: porting VERTa to Spanish to evaluate adequacy (Chapter 9). This allowed us to explore our initial hypothesis that when porting a linguistically-enhanced MT metric to a new language, the metric would have to consider the main characteristics of the new language. In our case, which linguistic features in our metric would have to be slightly modified, which changes would have to be performed and finally if the metric would be easy to adapt to a new language. To this aim the Spanish corpus was used (see section 3.2.1.5) and the linguistic features used were adapted to the main characteristics of Spanish to evaluate adequacy. Furthermore, VERTa was compared to other well-known metrics used to evaluate Spanish, showing that it also outperformed them.

## 3.2 Corpus Data, other Resources and Tools

In order to perform our research, several resources and tools were required to carry out our initial linguistic analysis, develop our MT metric and conduct experiments. Below, the data used to conduct the linguistic analysis, develop and evaluate our metric is described (section 3.2.1), as well as other resources and tools used in our metric (section 3.2.2).

### 3.2.1 Data

In our study several data sets have been used to perform the linguistic analysis, develop our metric and study the influence of the linguistic features used, conduct a meta-evaluation of the metric and finally port our metric to Spanish. These data come from three different sources:

- the MetricsMatr 2010 evaluation task[14], whose data was used to perform the linguistic analyses and experiments based on adequacy in English, as well as the meta-evaluation on adequacy,

[14] http://www.statmt.org/wmt10/evaluation-task.html

- the NIST 2005 Open Machine Translation (OpenMT) Evaluation campaign[15], whose data was used to conduct experiments based on fluency in English, as well as the meta-evaluation on fluency,
- the WMT12[16] and WMT13[17] Metrics Shared Tasks, whose data was used to prepare VERTa to participate in the WMT14 Metrics Shared Task,
- the WMT14 Metrics Shared Task[18], whose data was used to participate in the metrics shared task.
- and, finally, the KNOW2 project[19], whose data was used to port our MT metric to Spanish and evaluate adequacy.

All data sets contain hypothesis segments (the MT output evaluated), reference segments (human translations) against which the hypothesis segments are compared, and segment-level human judgements used to correlate scores provided by the metric in order to evaluate it, with the exception of data from WMT14, where human judgements were not provided.

Next, each of the data sets are described.

#### 3.2.1.1 MetricsMatr2010 Evaluation Task Data

The first corpus, provided by the MetricsMatr organizers, was part of the development data of the MetricsMatr evaluation task, which belongs to newswire genre and was divided into two parts: one part for the linguistic analysis and the development of the metric, whereas the second part was kept unseen in order to evaluate the metric. The first part of the corpus, which was used to conduct the linguistic analyses and develop VERTa (see Chapters 4 and 6) consisted of 100 segments (Arabic to English) of the NIST Open-MT06 data, the MT output from 8 different MT systems and 4 reference translations. The human judgments used were based on adequacy (7-point scale, adjudicated judgements). All segments were taken into account regardless of the system

[15] https://catalog.ldc.upenn.edu/LDC2010T14
[16] http://www.statmt.org/wmt12/metrics-task.html
[17] http://www.statmt.org/wmt13/metrics-task.html
[18] http://www.statmt.org/wmt14/metrics-task/
[19] http://www.ehu.es/ehusfera/know2/2011/02/08/intro/

providing them, in order to have a more precise correlation and avoid being system-biased.

The second part of the corpus, kept to evaluate the performance of the metric, contained 149 segments translated by 8 different systems, 4 reference translations and the corresponding adjusted human judgements for adequacy.

#### 3.2.1.2 NIST 2005 Open Machine Translation Evaluation Campaign Data

This data set was granted by NIST[20] and LDC[21] from their NIST 2005 Open Machine Translation (OpenMT) Evaluation campaign. Similar to the MetricsMatr data, this data was divided into two parts: one for the development of the metric and the other for its evaluation. The part aimed at the development of the metric included 100 segments from Arabic into English, the MT output from 6 different systems, 4 reference translations and level-segment human judgements by 2 judges on fluency, from which a final judgement was obtained by calculating the average.

The second part of the corpus used to evaluate the metric as regards fluency contained 149 segments, the MT output of 6 different systems, 4 reference translations and human judgements per segment provided by 2 different judges, which were averaged to obtain a single human judgement.

#### 3.2.1.3 WMT12 and WMT13 Metrics Shared Task Data

These data sets were provided by the WMT organisation from the WMT12 and WMT13 Metrics Shared Tasks. From these data sets, we selected and use all of those referring to all languages into English (en). Being "all languages" French (fr), German (de), Spanish (es) and Czech (cz) for WMT12 and French, German, Spanish, Czech and Russian (ru) for WMT13. Data sets distributed are reported in Table 2 and Table 3. These data sets were used to conduct initial experiments before participating in the WMT14 Metrics Shared Task (see section 8.3 for further details).

[20] http://www.nist.gov/
[21] https://www.ldc.upenn.edu/

| WMT12 Data | cs-en | de-en | fr-en | es-en | Total |
|---|---|---|---|---|---|
| **#systems** | 6 | 16 | 15 | 12 | 49 |
| **#segments per system** | 3,003 | 3,003 | 3,003 | 3,003 | 12,012 |
| **#segments** | 18,018 | 48,048 | 45,045 | 36,036 | **147,147** |

**Table 2** WMT12 data

| WMT13 Data | cs-en | de-en | fr-en | ru-en | es-en | Total |
|---|---|---|---|---|---|---|
| **#systems** | 12 | 23 | 19 | 23 | 13 | 90 |
| **#segments per system** | 3,000 | 3,000 | 3,000 | 3,000 | 3,000 | 15,000 |
| **#segments** | 36,000 | 69,000 | 57,000 | 69,000 | 39,000 | **270,000** |

**Table 3** WMT13 data

In both campaigns only one reference was used and the evaluation based on ranking was conducted at system and segment level.

#### 3.2.1.4 WMT14 Metrics Shared Task Data

This data was provided by the WMT14 organisation and contained MT output from "all languages" into English, being all languages Czech, German, Hindi (hi), French and Russian (details on the data provided are reported in Table 4), and 1 reference translation was used. In this case, human judgements were not provided since they were later used by the organizers to decide the position of the participating metrics in the shared task.

| | cs-en | de-en | hi-en | fr-en | ru-en | Total |
|---|---|---|---|---|---|---|
| **#systems** | 5 | 13 | 9 | 8 | 13 | 48 |
| **#segments per syst.** | 3,003 | 3,003 | 3,003 | 3,003 | 3,003 | 15,015 |
| **#segments** | 15,015 | 39,039 | 27,027 | 24,024 | 39,039 | **144,144** |

**Table 4** WMT14 data

This data was used to participate in the WMT14 Metrics Shared Task, which examines MT evaluation metrics with the aim of achieving the strongest correlation with human judgements of translation quality (please, refer to section 8.3 for further details).

#### 3.2.1.5 KNOW2 Project Data

The last data set belongs to the glosses of the Spanish WordNet developed in the KNOW2 project and it is smaller than the English data set. This data was used in order to work on the portability of VERTa to Spanish. The data contains: 187 WordNet glosses that were translated from English into Spanish by means of two different systems (Apertium[22] and Google Translator[23]); four reference translations, one extracted from Spanish WordNet 1.6 and the rest produced by human translators; and human judgements on adequacy provided by two different judges, whose average was calculated to obtain a single human judgement.

### 3.2.2 Other Resources and Tools

Apart from the data described above, there were other resources and NLP tools that are used in the VERTa metric. These resources are language-dependent, since they depend on the language evaluated, in our case English and Spanish. Next, resources have been organised and listed depending on the language evaluated and the module inside the metric where they are used.

#### 3.2.2.1 Resources for English

Those resources and tools[24] used in the English VERTa in a per-module basis are:

- Lexical Module.
    - WordNet 3.0 (Fellbaum 1998). WordNet 3.0 is a large lexical database that contains nouns, adjectives, verbs and adverbs organised into synsets (cognitive synonyms), each expressing a distinct concept. Synsets are interconnected by means of conceptual semantics and lexical relations. In VERTa, WordNet has been used in order to obtain information regarding synonyms, hypernyms and hyponyms and also to lemmatize the corpus. As regards hypernyms and hyponyms, the most frequent word sense is used. In addition, in the Princeton WordNet there is a library that

[22] http://www.apertium.org/
[23] http://translate.google.com/
[24] Further information on the resources and tools used is provided in Appendix A.

contains a lemmatizer which has been used to obtain the lemmas corresponding to word-forms in our data.

- Morphological Module.
    - The Standford Log-Linear Part-of-Speech Tagger (Toutanova et al. 2003), included in the the Standford CoreNLP suite, has been used to PoS tag our corpus.
- Dependency Module.
    - The PCFG parser for English (KLein and Manning 2003; de Marneffe et al. 2006) contained in the Standford CoreNLP suite has been used to obtain the dependency relations in our corpus.
- Semantic Module.
    - Named Entity Recognition (NER). In order to identify NEs the Supersense Tagger (Ciaramita and Altun 2006) has been used.
    - Named Entity Linking (NEL). The NEL metric uses a graph-based NEL tool inspired by Hachey et al. (2011), which links NEs in a text with those in Wikipedia pages.
    - Time Expressions recognition and normalization. The Stanford Temporal Tagger (Chang and Manning 2012), contained in the Standford NLP suite, has been used, which recognizes not only points in time but also duration.
    - Sentiment analysis. The dictionary strategy described in Atserias et al. (2012) has been used in order to compute the contextual polarity of a segment. In other words, it uses a dictionary strategy to determine whether the contextual polarity of a sentence is negative, neutral or positive.

- Language Model Module.
    - News LM[25]. This LM was used as a baseline feature in the WMT13 Quality Estimation Task[26]. This resource was built from the news data released as part of WMT11 and in the present work it has been used for widening the coverage of those segments that, even being syntactically different from their corresponding reference translations, are still fluent (see section 7.7 for further details).

#### 3.2.2.2 Resources for Spanish

The resources and tools used in the Spanish VERTa are:

- Lexical Module.
    - WordNet 3.0 (Fellbaum 1998) for Spanish has been used in order to obtain information regarding synonyms, hypernyms and hyponyms.
    - Freeling (Padró and Stanilovsky 2012) has been used to lemmatize our corpus.
- Morphological Module.
    - The PoS Tagger Module in Freeling has been used to PoS tag our corpus.
- Dependency Module.
    - The Txala parser (Atserias et al. 2005), with the dependency grammar developed for Spanish (Lloberes et al. 2010) included in Freeling, has been used to obtain the dependency relations in our Spanish corpus.

[25] http://www.quest.dcs.shef.ac.uk/quest_files/de-en/news.3gram.en.lm
[26] http://www.statmt.org/wmt13/quality-estimation-task.html

### 3.3 Summing Up

This chapter has covered the methodology followed to perform the research presented in this work. The approach taken has been an empirical one, based on data analysis and experiments to formulate and test our hypothesis.

In section 3.1 all the steps followed to conduct our research have been detailed: firstly, the linguistic analysis conducted and how this analysis was the focus for the development of VERTa, our MT metric and tool used to perform our experiments; secondly, the experiments carried out by means of VERTa, which allowed us to perform both a quantitative and qualitative analysis on the suitability and use of those linguistic features tested, looking at the quantitative part as a complement to our linguistic gist; thirdly, the meta-evaluation stage, where VERTa was compared to other well-known MT metrics to show that it was not only a tool to do our experiments, but it could be applied as a state-of-the-art MT metric; finally, porting VERTa to Spanish so as to test the linguistic information that had to be modified to evaluate Spanish output on adequacy and thus if the metric was easy to port. Even if not necessarily one of the main objectives of this thesis, VERTa demonstrates its ability to work as an MT evaluation metric and to be ported to other languages.

Finally, all data, resources and tools used in the present study are detailed in section 3.2. First, the data sets used to perform the linguistic analysis, experiments and meta-evaluation are described (section 3.2.1), then the resources and tools used, either for data processing or in the metric's architecture, are listed in section 3.2.2.

The following chapter is devoted to the initial linguistic analysis conducted with the data described in section 3.2.1.1.

# Chapter 4. Linguistic Analysis of Data

Several linguistic phenomena must be analysed and considered when approaching MT evaluation from a linguistic point of view, especially when using reference translations for the purpose of comparison. In the last decade there have been several studies of MT errors, the most well known being Vilar et al. (2006) and Farrús et al. (2010). The focus of these studies, though, is rather narrow as they are only based on the output obtained from statistically-based MT systems and they just highlight the errors made by a single MT system, the system under study in each case. Our study widens this approach in three ways: 1) in order to cover a wider range of linguistic phenomena our study is not restricted to a single system; on the contrary, it analyses the MT output of several systems; 2) opposite to the studies previously mentioned, which only focus on statistically-based systems, the data analysed in this paper covers both rule-based and statistically-motivated MT systems; 3) the analysis conducted is not only restricted to the identification of MT errors but it also identifies those positive linguistic points that an MT evaluation metric must take into account, by means of comparing MT output to reference translations.

The approach followed by our analysis was initially a corpus-based approach (Tognini-Bonelli 2001), because there were some linguistic phenomena that we already expected to find in our data. However, while the analysis was being carried out, some new phenomena arose, leading us to a more corpus-driven approach.

As regards the classification of linguistic phenomena, Vilar et al. (2006) provided a very detailed classification of translation errors, as shown in Figure 6[27]. However, this classification is rather vague and does not follow any linguistic criteria (e.g. the category "Incorrect words" covers words kept in the source language, mistranslated words due to the ambiguity of meaning of the source word, mistranslated words due to its morphological form). Farrús et al. (2010), though, opted for a more general and linguistically-oriented classification organising MT errors into different linguistic levels: orthographic errors, morphological errors, lexical errors, semantic errors and syntactic errors. This classification covering the different levels of language is more

[27] Figure taken from Vilar et al. 2006.

appropriate to our needs, mainly because it is a wide and language-independent classification which allows us to deal not only with errors but also with those positive characteristics that must be considered.

In the following, the linguistic characteristics analysed and their classification is presented. Section 4.1 describes the data used for our analysis briefly; section 4.2 focuses on the linguistic analyses carried out, covering format and orthography (section 4.2.1), the lexical level (section 4.2.2), morphological level (section 4.2.3), syntactic level (section 4.2.4) and semantic level (section 4.2.5); finally, a summary and the findings of this chapter are presented in section 4.3.

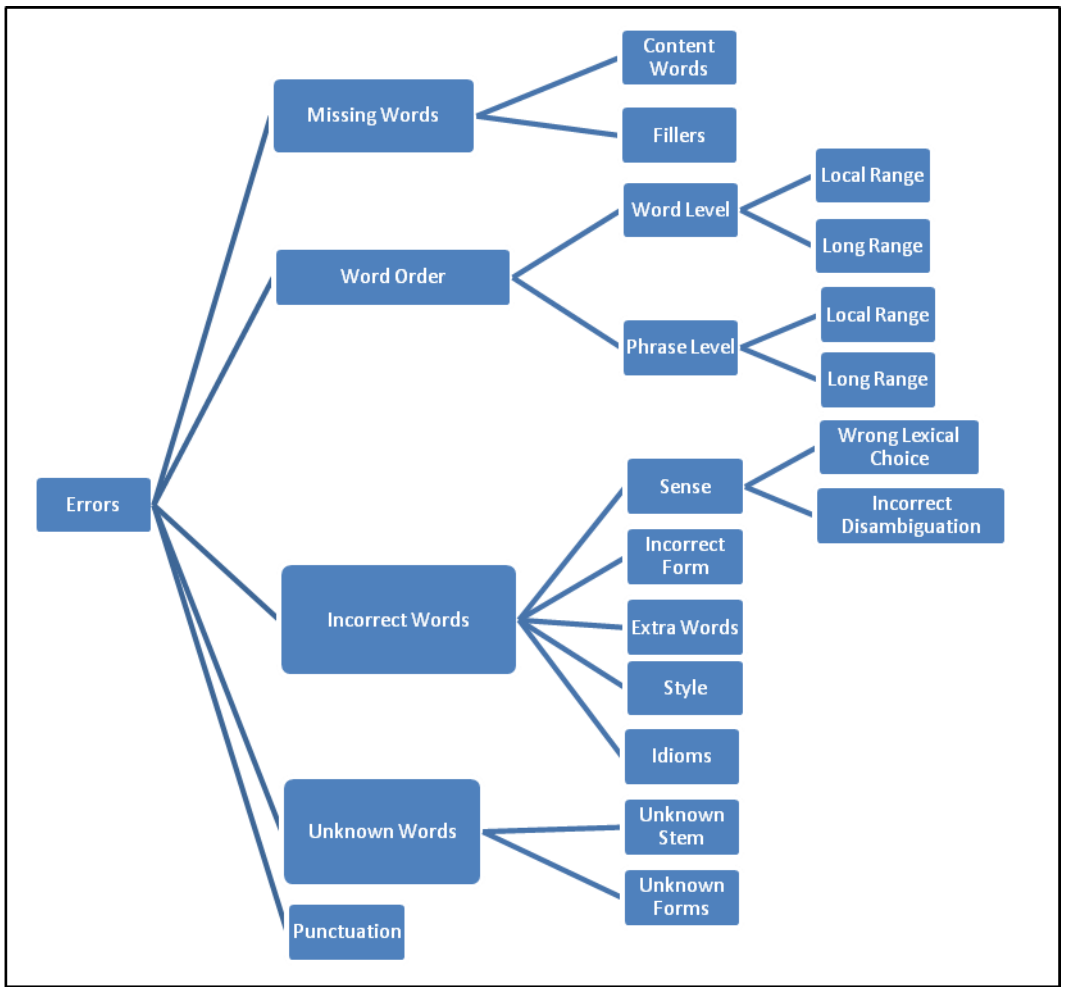



**Figure 6** Classification of Translation Errors (Vilar et al. 2006)

## 4.1 Data Analysed

Regarding the data used in this study, as described in Chapter 3, it comes from two different sources: the MetricsMatr 2010 evaluation task[28] and the KNOW2 project[29]. The first corpus is part of the development data of the MetricsMatr evaluation task that belongs to newswire genre. From this development corpus 100 MT output segments (Arabic to English) and 400 segments corresponding to 4 different reference translations were chosen to perform the initial linguistic analysis. Thus, the data analysed corresponds to the comparison between the MT output segments in English and the reference translations, also in English. The rest of the data from the development corpus was kept in order to perform experiments and conduct the quantitative and qualitative analysis of the results obtained (see Chapters 6 and 7). The second corpus belongs to the glosses of the Spanish WordNet developed in the KNOW2 project and it is smaller than the English corpus. It consists of 187 MT output segments from English into Spanish and a total of 748 Spanish reference translations. Since this is such a small corpus we use the whole set for both the initial linguistic analysis and the experiments (see Chapter 8).

## 4.2 Linguistic Analysis

In the linguistic analysis shown next, linguistic features have been organised as follows:

1. Format and orthography, including punctuation marks, different date, time and money formats and letter capitalisation;

2. Lexical level, including non-translated source words, missing target words, no semantic correspondence between source and target words, non-translated proper nouns and those translated when not necessary or wrongly translated;

3. Morphology, including inflectional and derivational morphology, compounding and morphosyntax;

4. Syntax at phrase and clause level, including word order, alternations, wrong prepositions and ungrammatical chunks;

[28] http://www.statmt.org/wmt10/evaluation-task.html
[29] http://www.ehu.es/ehusfera/know2/2011/02/08/intro/

5. Semantics, including lexical semantic relations (i.e. synonymy, hypernymy, hyponymy, polysemy and homonymy) and sentence semantics (i.e. non-translated semantic arguments).

Although some of these parts are inherently related, they have been approached separately for the sake of analysis.

### 4.2.1 Format and Orthography

This section addresses several linguistic issues related to format and orthography: the use of capital letters, date and time formats, the realisation of quantities either by means of number or letters, the use of symbols and finally, the use of punctuation. Although it can be argued that format is not strictly a linguistic issue, it has been included in this section because it deals with the form of lexical items and different possible linguistic realisations of certain expressions.

**Capital letters.** The use of capital letters in both hypothesis (HYP) and reference (REF) segments is first analysed.

- In Arabic, one of the source languages in our data, there is no use of capital letters, causing that in many English hypothesis translations the word following a full stop does not start with a capital letter, as shown in Example 1.

  Example 1

  HYP: *A delegation from Hamas Monday started talks in Cairo on the formation of the new Palestinian government.* ***after*** *having held the leadership of the Movement meetings …*

  REF: *On Monday a Hamas delegation began talks in Cairo regarding the formation of the new Palestinian government.* ***After*** *the leaders of the movement held two meetings…*

- In addition, proper names are also affected since some words such as titles are written in upper or lower case. It has been noticed that the use of upper case is usually preferred in the reference string whereas in the hypothesis translation,

MT engines tend to skip them. For instance, in the data analysed the word *prophet* and *Prophet* were found, as exemplified below.

Example 2

HYP: ...***prophet*** *Mohammed...*

REF: ... ***Prophet*** *Mohammed...*

**Date and time format.** Regarding date and time format, there are several ways to express them that need to be tackled when comparing the hypothesis and reference strings. Dates can be realised by means of numerical expressions or letters (e.g. *6-2*, *February 6*, *February 6th*), by means of different signs (e.g. *2-6* or *2/06*), in a reversed mode (e.g. *6-2*, European way, or *2-6*, American way), etc. In the data analysed the equivalent expressions shown in Example 3 were found. These temporal expressions are semantically identical but have a different surface realisation, therefore they should not be penalised when comparing hypothesis and reference segments, but regarded as possible equivalents.

Example 3

HYP: *Oslo* ***6-2*** *(AFP)...*

REF1: *Oslo* ***2-6*** *(AFP)...*

REF2: *Oslo,* ***February 6*** *(A.F.P.)...*

REF3: *Oslo* ***2/06*** *(A.F.P.)...*

**Quantities and currency.** Another issue under consideration is the way quantities are expressed, either by using numbers, letters or a combination of both.

- As illustrated in the example below, in the hypothesis string there is a mixture of numbers (*20*) and letters (*thousand*). In the first reference translation the use of the numerical expression is preferred (*20,000*), whereas in the second reference segment the use of letters is favoured (*twenty thousand).*

Example 4

HYP: ...***20 thousand*** *dollars...*

REF1: ...***20,000*** *dollars...*

REF2: ...***twenty thousand*** *dollars...*

- A similar case is that of currency which can also be expressed by means of letters or symbols as shown in Example 5.

Example 5

HYP: ...*20 thousand* ***dollars****...*

REF: ...**$** 20,000...

**Punctuation marks.** One more point that also deserves especial attention is the use of punctuation marks.

- This issue is especially important when dealing with acronymy, because sometimes each letter is separated by a punctuation mark whereas other times it is not. Although some punctuation standards recommend not to use periods in acronyms or abbreviations based on initial letters, this recommendation is not always followed. Such an example is found in the acronymy for the Agence France Press in the following segments:

Example 6

HYP: *Oslo 6-2 (****AFP****)...*

REF: *Oslo, February 6 (****A.F.P.****)...*

In the hypothesis segment the acronym for Agence France Press is written without punctuation marks (*AFP*) in the hypothesis, whereas in the reference sentence periods are being used (A.*F.P.*).

- Punctuation marks must also be under consideration regarding the order of appearance inside the sentence, which does not always coincide when reference

and hypothesis are compared. For instance, when finishing a sentence with a quotation, as illustrated in the example below, the order of the punctuation can vary. In the hypothesis string we first find inverted commas followed by full stop, whereas in the reference translation the full stop is followed by inverted commas.

Example 7

HYP: *...to set fire***".**

REF: *...lighting fires.***"**

- Besides, sometimes punctuation marks are omitted by the MT engine, therefore affecting the fluency of the sentence. An example is found in the Spanish data analysed as regards the use of question marks, as shown in the example below. In correct written Spanish question marks must appear both at the beginning and at the end of the clause, as exemplified in the reference segment. However, MT engines may avoid the use of the question mark at the beginning of the clause, due to the source language (i.e. English), resulting in a drawback to the fluency of the sentence.

Example 8

SOURCE: *is the direction of the economy a function of government?*

HYP: ***ø*** *es la dirección de la economía una función de gobierno?*

REF: ***¿****es la dirección de la economía una función de gobierno?*

Up to now format and punctuation issues have been under consideration, from now on we move to a more linguistic level and we start dealing with lexical units.

### 4.2.2 Lexical Level

In this section we focus on the elements at lexical level. There are several points that deserve especial attention, from the role of multi-word units in MT evaluation to non-translated or wrongly translated proper nouns, without forgetting the use of acronymy and abbreviation.

**Multi-word units**. Multi-word expressions are a problem for Natural Language Processing (NLP) in general, and for automatic MT evaluation, in particular, mainly due to their format and their meaning. In NLP, multi-word expressions are usually defined according to Sag et al.'s definition: "*idiosyncratic interpretations that cross word boundaries (or spaces)*" (Sag et al. 2002:2). In our study, under the term multi-word unit we will consider endocentric, exocentric and copulative compounds ( *blockhead* or *Arab* Israeli conflict), verb particle constructions (*go on*), idioms (*stab in the back*) and fixed expressions (*with regard to*).

- Regarding their format, they present a similar problem to those covered in the formatting section, in the sense that multi-word expressions can sometimes be written with a hyphen, without a hyphen, or separated by a space, such as the unit *Arab Israeli* in Example 9, which in the hypothesis segment is written separated by a space whereas in the reference segment it is joined by a hyphen.

  Example 9

  HYP: ... *the* ***Arab Israeli*** *conflict...*

  REF: ... *the* ***Arab-Israeli*** *dispute...*

- As for their meaning, it also turns into a problem when the meaning of the multi-word expression can be expressed by a single word since NLP tools may have difficulties to establish the relationship between one single lexical unit and more than one. That is usually the case with verb particle constructions, idioms and fixed expressions, as illustrated in the examples below where the multi-word expression *with regard to* in the reference translation is expressed by the single lexical unit *concerning* in the hypothesis string.

  Example 10

  HYP: ...***concerning*** *its nuclear power...*

  REF: ...***with regard to*** *its nuclear power...*

Also concerning the meaning of multi-word expressions, non-decomposable compounds and fixed expressions must be taken into account. Some MT engines

translate these lexical structures word for word, resulting in a wrong (and sometimes weird) translation. Example 11 illustrates a wrong translation of the English idiom *cut someone off at the knees* into Spanish.

Example 11

SOURCE: *Obama should cut him off at the knees*

HYP: *Obama debe cortar él por las rodillas*

REFERENCE: *Obama debería pararle los pies*

The meaning of *cut him off at the knees* is opaque, in other words, it cannot be translated word for word. However, the MT engine used to translate this idiom has not been able to get the Spanish equivalent *pararle los pies*, instead a literal translation was provided. As a consequence, the hypothesis segment is nonsense and fails in translating the meaning of the source text.

**Acronyms and abbreviations.** Acronyms are words "*formed by combining the initial letters of the principal words in a phrase*" (Trask 1992:5) whereas abbreviations are, according to the Macmillan English Dictionary "*a short form of a word or a phrase*". This means that the extended form of the acronym or abbreviation is made up of one (in the case of the abbreviation), two or more lexical items. The implicit relation existing between the abbreviation and its expanded form and, especially between the acronym and its extended form needs also to be taken under consideration as shown in the following example:

Example 12

HYP: *Birmingham (**US**)...*

REF: *Birmingham (**United States**)...*

The acronym *US* in the hypothesis relates directly to its extended form *United States* which appears in the reference translation.

So far, we have covered the relationship between a semantic referent and its surface form, as regards acronyms and abreviations, quantities, currency and multiwords. Next

we focus on lexical content and we will cover non-translated source words, missing target words and we will finish this subsection by dealing with proper nouns and their lack of translation or mistranslation.

**Non-translated source words.** Some MT engines, especially rule-based MT engines are not able to translate all words in the source sentence because they may not be part of their lexicons. Therefore when these engines find an out-of-vocabulary word, they opt for leaving the source word untranslated in the target translation (see Example 13). It is obvious that the presence of a source word in the MT output will affect both adequacy and fluency of the sentence, although it may vary depending on the word and the quantity of words. In Example 13 the out-of-vocabulary word *rumiant* poses a minor drawback to the understanding of a sentence since both source and target words are quite similar and a potential Spanish user might be able to understand the meaning of a sentence. Unfortunately, that is not the case in Example 14, where 2 out of 3 words have been kept in the source language preventing the user from understanding the whole segment.

Example 13

SOURCE: *the second compartment of the stomach of a* ***rumiant.***

HYP: *El segundo compartimento del estómago de un* ***rumiant****.*

REF: *El segundo compartimento del estómago de un* ***rumiante****.*

Example 14

SOURCE: ***robbery*** *at* ***gunpoint***

HYP: ***Robbery*** *en* ***gunpoint****.*

REF: ***robo*** *a* ***punta de pistola***.

**Missing target words and proper nouns.** Another issue regarding the lexical level is missing target words. These words can be either content words and therefore the meaning of the sentence is highly affected, or function words which will cause a stronger effect on the grammaticality of the sentence. Such is the case of the following

example, where the English sentence has been translated word for word into Spanish, triggering the omission of definite article *la* between the preposition *de* and the noun *comida* in the candidate translation; thus resulting into a disfluent segment.

Example 15

SOURCE: *he took a walk after lunch*

HYP: *tomó un paseo después de ø*[30] *comida*

REF: *dio un paseo después de* ***la*** *comida*

However, when the missing part in the hypothesis string is a content word or a proper noun the loss of meaning becomes really important. Example 16 illustrates how the proper noun *Valentine's* has not been translated, resulting in a loss of meaning. It must also be highlighted that, as stated in the introduction to this linguistic analysis (see section 4.2), although linguistic phenomena are presented separately for the sake of analysis, it is very common to find instances of more than one phenomenon in a single string, such as the omission of the function word *a* also in Example 16.

Example 16

HYP: *The ø day is ø very popular day in Iraq...*

REF: ***Valentine's*** *day is a very popular day in Iraq...*

At its turn Example 17 evidences the loss of meaning when a content word is not translated, as illustrated by the untranslated verb *hold* in the hypothesis segment, which clearly impedes the understanding of the sentence.

Example 17

HYP: *It is noteworthy that the Iraqi President 's son Oudi was known to ø a very large ceremonies...*

[30] *ø* indicates omission of a target word.

REF: *It is noteworthy that the son of the Iraqi President Uday was known to* ***hold*** *huge parties...*

In the data under study, determiners, punctuation marks, proper nouns and existential THERE + BE are those which tend to disappear in the hypothesis string.

Example 18

HYP: *The victims...*

REF: ***There were no*** *victims...*

However, not all words that are omitted can cause a problem in the fluency or adequacy of the segment. There are some words which can be omitted without altering the meaning or comprehensibility of a sentence, for instance the optional conjunction *that* which can be omitted in English under certain circumstances.

Example 19

HYP: *...He said "I believe* ***that*** *the situation..."*

REF: *...He added "I think* ***ø*** *the situation..."*

**Wrong translation of proper nouns.** In the previous section the lack of translation of proper nouns was covered. However, this is not the only issue with proper nouns as it is also common for MT engines to mistranslate them, as shown in the example below, where the proper noun *Zeev Boim* is partly missing (i.e. *Boim*) and partly mistranslated (i.e. *Zaif* vs. *Zeev*)

Example 20

HYP: *...****Zaif*** *told reporters...*

REF: *...****Zeev Boim*** *said in a press statement...*

Once those issues that must be covered at lexical level have been determined, we move now to morphology, which by means of morphosyntax will help us connect the lexical level to the syntactic one.

### 4.2.3 Morphological Level

This section covers both inflectional and derivational morphology, as well as morphosyntactic features. According to Katamba, morphology is the "*analysis of the internal structure of words*" (Katamba 1993:1). It can be classified into derivational and inflectional morphology. The former refers to the creation of new lexemes by means of affixation and compounding, whereas the latter deals with the creation of new word-forms belonging to the same lexeme.

**Derivational morphology.** When speaking about derivational morphology in MT evaluation, a couple of word formation processes must be analyzed: affixation and compounding.

- In the case of affixation, words that contain the same root but belong to a different lexeme are considered, since the root contains the core meaning of a word. That is the case in Example 21 where the words *participation* in the candidate translation and *participate* in the reference share the same root *particip-*. Although the former is a noun and the latter a verb and they occur in different syntactic structures, the fact of sharing the root helps in making the structures semantically similar (see subsection on Syntactic Structures in section 4.2.4).

  Example 21

  HYP: *...to open **participation** in government...*

  REF: *...to **participate** in government...*

**Inflectional morphology.** This is an important element, especially when dealing with languages with rich inflectional morphology, such as Spanish, French, Catalan, etc. because it helps to deal with linguistic features such as tense, aspect, mood, number, gender or case.

- By means of morphological features we can compare whether the word-form used in the hypothesis and the reference translation is the same or varies (i.e. in the case of tense, whether both segments refer to a present or past action), as

exemplified below by means of the *–ed* ending in the verb form *carri**ed*** which provides information on the tense of the verb, in this case past tense in both hypothesis and reference.

Example 22

HYP: *...dozens of suicide attacks **carried** out by Hamas...*

REF: *... dozens of suicide attacks **carried** out by Hamas..*

- On the other hand, wrong translations due to non-equivalent verb tense between source and target strings can also be identified by means of morphology. The analysis of Example 23 shows that English past simple *was* is translated into the Spanish past *era* instead of *fue*. Although the verb is correctly selected, as both refer to the verb TO BE, and both refer to the past, the problem lies on the verb tense. The verb form *era* in the hypothesis sentence is in the imperfect tense, showing that the action is not completed, whereas in the reference translation *fue* is a preterite, conveying the meaning that the action expressed by the verb is completed.

Example 23

SOURCE: *his success in the marathon **was** unexpected*

HYP: *su éxito en el marathon **era** inesperado*

REF: *su éxito en el maratón **fue** inesperado*

- Similarly, mistranslations of the English nominalised –ing form into the Spanish gerund are also found in our data. In English, the –ing form can refer to an infinitive, to a gerund or even to a noun or adjective. This –ing form is sometimes mistranslated into Spanish, affecting the fluency of the sentence, as shown in the following example:

Example 24:

HYP: *en el **pulsando** de un botón*

REF: *al **pulsar** un botón*

The Spanish gerund *pulsando* is a wrong translation of the nominalised –ing English form *at the pressing of* which should be translated as the Spanish infinitive *pulsar*.

- Besides, verb forms are not always translated correctly. Example 25 shows that the third person singular past form of *do* has been translated as an infinitive form *hacer* instead of the finite verb form *hizo*.

  Example 25:

  HYP: *Él no **hacer** un movimiento para ayudar*.

  REF: *Él no **hizo** un movimiento para ayudar.*

**Morphosyntactic features.** Morphosyntactic features are "*properties that are partly morphological and partly syntactic*" (Katamba 1993:14). Inflectional morphology in combination with syntax (morphosyntax) also plays a key role in the sentence fluency. Such is the case of agreement in English, where verb forms in third person singular show agreement with the subject by means of the *–s* ending. In other languages with richer inflectional morphology such as Spanish, morphosyntax plays even a more important role not only in the agreement between subject and verb regarding person and number (e.g. ***Yo como** poco* – "I eat little" vs. ***Ellos comen** poco* – "They eat little") or subject and participle regarding person, number and gender (e.g. *las **casas** fueron **diseñadas –*** "the houses were designed" vs. *el **edificio** fue **diseñado*** – "the building was designed"); but also inside the noun phrase between determiner, noun and adjective, regarding gender and number (e.g. *un**a** chic**a** alt**a*** – "a tall girl" vs. *un**os** chic**os** alt**os*** – "some tall boys"); and between the subject and the subject complement (e.g. *mis hij**as** son rubi**as*** – "my daughters are blonde").

After dealing with morphosyntax, in the following section, we move from morphology to our next level under analysis, the syntactic level.

### 4.2.4 Syntactic Level

At this level several issues are considered: a) wrong translations affecting syntax, such as errors when translating prepositions, or errors when translating relative clauses and reflexive pronouns; b) the syntactic structure; and c) word order both inside the phrase and inside the clause. Under syntactic structure we cover those changes that imply a change of grammatical category, whereas those structures included in word order do not entail a change in the grammatical category of the units affected.

**Incorrect translations.** As for incorrect translations affecting syntax we identify a couple of items: first, wrong prepositions and later, errors in relative clauses.

- Sometimes prepositions are not translated correctly, thus influencing both the adequacy and fluency of the sentence. In our Spanish corpus the following example was found:

  Example 26:

  SOURCE: *The act of appearing **in** public view…*

  HYP: *El acto de aparecer **en** la vista del público...*

  REF: *El acto de aparecer **a** la vista del público...*

  The preposition in bold *en* is a mistranslation of the English preposition *in* in the chunk *in public view*. In this case although the meaning of the sentence is not seriously damaged, fluency is. In fact, a native speaker of the language would never use such a preposition in that context.

- Regarding relative clauses, some MT engines have problems when translating relative constructions as shown in Examples 27 and 28 where relative constructions *from which* and *whose,* respectively, have been mistranslated or not translated at all. In Example 27 there is a morphological issue related to gender agreement between the relative construction and its referent; whereas in Example 28, the MT engine has failed in translating the relative construction.

Example 27

SOURCE: *installation **from which** a military force*

HYP: *instalación **del cual**....*

REF: *instalación **de la cual**...*

Example 28

SOURCE: *a car driven by a person **whose** job is....*

HYP: *un coche conducido por una persona **whose el** trabajo es...*

REF: *un coche conducido por una persona **cuyo** trabajo es...*

- In relation to reflexive pronouns, in Spanish reflexive actions are expressed by means of reflexive pronouns attached to the same verb. In the example below, verb *matar* in the hypothesis sentence does not contain a relative pronoun although from a syntactic point of view it is required, as shown by *matarte* in the reference.

  Example 29

  SOURCE: *The act of **killing yourself***

  HYP: *El acto de **matar a ti mismo***

  REF: *El acto de **matarte a ti mismo***

**Syntactic structure.** As regards the syntactic structure, it is possible to find expressions which are similar in meaning and with very close lexical items but showing different syntactic structures and grammatical categories, as shown in the example below.

Example 30

HYP: *...Putin **invited** Hamas...*

REF: *...Putin**'s invitation to** Hamas...*

From a semantic point of view, both the hypothesis and reference segments express the same meaning and the lexical items used are also very close – *invited* and *invitation* share the same root *invite*, however, the syntactic structures are quite different. The

hypothesis string is realised by a clause (see Figure 7) – where *Putin* is the Subject, *invited* the verb and *Hamas* the direct object, whereas the reference string is a Noun Phrase (NP) (see Figure 8) whose head is the noun *invitation* with a genitive pre-modifier *Putin's* and a Prepositional Phrase (PP) *to Hamas* working as a postmodifier of the noun *invitation*. Thus, although performing a different syntactic structure both hypothesis and reference segments are semantically equivalent.

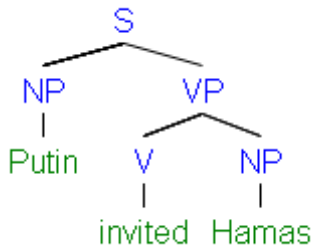



**Figure 7** Tree diagram corresponding to the hypothesis string

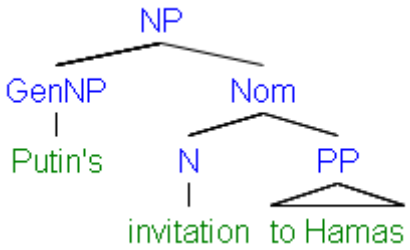



**Figure 8** Tree diagram corresponding to the reference string

This change of syntactic structures is very common in English Noun Phrases where the head of the phrase is postmodified by a Prepositional Phrase introduced by the preposition *of* (see Figure 9). This postmodifier is shifted to a position preceding the noun by removing the preposition *of* and working as a NP pre-modifying the head (see Figure 10), as illustrated below.

Example 31

HYP: *...a delegation* ***of Moroccan police****...*

REF: *...a* ***Moroccan police*** *delegation...*

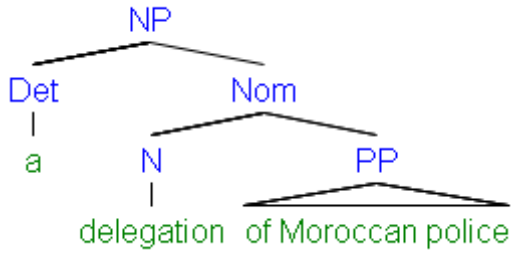



**Figure 9** Tree diagram corresponding to the hypothesis string

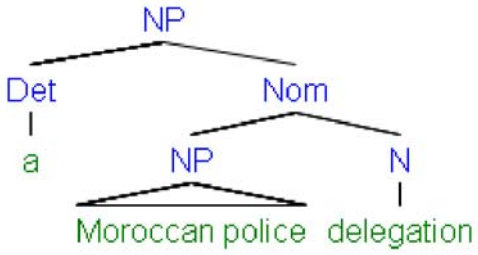



**Figure 10** Tree diagram corresponding to the reference string

At a clause level, one of the basic syntactic alternations is the use of passive and active voice; that is to say, by means of the active-passive alternation two different syntactic structures convey the same meaning. Example 32 shows that the verb in the hypothesis sentence is in the passive voice *were assassinated* and in the reference its voice is active *assassinated*. In the case of the agent, the entity performing the action expressed by the verb (Fillmore 1968), this appears in the form of a prepositional phrase, *by unknown men*, working as an Oblique in the hypothesis string; whereas in the reference string it occupies the position of a noun phrase, *unknown men*, working as the subject of the sentence.

Example 32

HYP: ...*were assassinated by unknown men*...

REF: ...*unknown men assassinated*...

Another valency alternation that must be highlighted is the dative/ditransitive alternation. In this case, the verb does not change at all but the two objects in the predicate exchange positions and phrase category, involving the addition or deletion of

a couple of prepositions (i.e. *to* and *for*) to the NP realising the indirect object. In the examples below, a dative alternation involving a to-PP and a for-PP is illustrated.

Example 33

HYP: *I fetched* ***Mary water***

REF: *I fetched* ***water for Mary***

Example 34

HYP: *I gave* ***him a present***

REF: *I gave* ***a present to him***

On the other hand, sometimes the hypothesis and reference segments are semantically equivalent but show very different vocabulary and quite different syntactic structure. This might be the case when the reference translation is a rather free translation, as shown in Example 35.

Example 35

SOURCE: *a narrow street with walls on both sides*

HYP: *Una calle estrecha con paredes en ambos lados*

REF: *Calle estrecha y larga entre paredes* ("*Narrow and long street between walls"*)

This is one of the most difficult cases to deal with in MT evaluation since the hypothesis segment is semantically and grammatically correct but it is very different from the reference segment.

**Word order.** In relation to word order, we must also distinguish word order at phrase level and word order at clause level.

- At phrase level, this phenomenon refers to those lexical items whose position inside the phrase changes either in the hypothesis or the reference string. In English this is very common in Noun Phrases containing more than one premodifier. As illustrated in the following example, the two adjectives

*Moroccan* and *official* premodifying the noun *source* occupy different positions in the hypothesis and the reference strings.

Example 36

HYP: *...a **Moroccan official** source...*

REF: *... an **official Moroccan** source...*

In the example above there is a slight change of meaning which does not affect at all the understanding of the chunk. However, in other cases the change of word order can affect the meaning in a deeper way, as illustrated below.

Example 37

HYP: *Nasser Abu Baker, **by***

REF: ***By** Nasser Abu Baker*

Example 38

SOURCE: *A formal **written** statement **of relinquishment***

HYP: *Una declaración formal **por escrito de la cesión***

REF: *Una declaración formal **de la cesión por escrito***

In Example 37, the preposition *by* appears at the end of the PP, which is absolutely ungrammatical in English and prevents the reader from fully understanding the sentence. As for Example 38, although both phrases are PPs, there is a problem in their word order, the one introduced by preposition *de* should precede the PP introduced by *por*. This is a consequence of a failure in reordering: in the English source sentence *written* modifies the whole NP *statement of relinquishment*, however it has been translated as if it only modified the noun *statement*.

- In relation to word order at sentence level, it refers to the order of the syntactic constituents inside the sentence. There are six types of possible word order depending on the language under analysis: SOV (e.g. Japanese and Turkish),

SVO (e.g. English and Romance languages), VSO (e.g. Classical Arabic), VOS (e.g. Fijian), OVS (e.g. Hixkaryana), OSV (e.g. Xavante). This issue becomes really important especially when dealing with a couple of languages with a completely different constituent word order, for example when translating from Arabic into English because MT systems tend to have problems with reordering constituents. Therefore, when developing an MT evaluation metric, word order of sentence constituents must be taken into account in order to capture this phenomenon. In the example below the subject *the concerned minister* occupies the position of the Object after the main verb *appeared* in the hypothesis sentence, whereas in the reference sentence the noun phrase is located before the verb, occupying the canonical position of the subject in English.

Example 39

HYP: *appeared* ***the concerned minister*** *saying....*

REF: ***The Minister concerned*** *replied saying...*

- In addition, it must be highlighted that some languages have a more flexible word order than others. This is plausible when dealing with a couple of languages such as English and Spanish. English word order is much stricter than Spanish word order. If we consider adjuncts, English adjuncts tend to appear either at the beginning of a sentence or at the end. Only some adjuncts of manner or degree can also appear between the subject and the verb, but no adjuncts can appear between the verb and the core complements as illustrated by the position of the adverb *yesterday* in the examples below, which causes one of the sentences (marked with an asterisk) to become ungrammatical.

  Example 40

  *(i)* ***Yesterday*** *the kids ate a lot of chocolate.*

  *(ii) The kids ate a lot of chocolate* ***yesterday****.*

  *(iii) *The kids ate* ***yesterday*** *a lot of chocolate.*

However, in Spanish the order of adjuncts is freer and they can appear almost in all places inside the sentence, as shown by the position of the adverb *ayer* ('yesterday' in English) in the following examples:

Example 41

*(i)* ***Ayer*** *los niños comieron mucho chocolate.*

*(ii)* *Los niños comieron mucho chocolate* ***ayer****.*

*(iii) Los niños comieron* ***ayer*** *mucho chocolate.*

*(iv) Los niños* ***ayer*** *comieron mucho chocolate.*

Therefore, it is worth taking into account constituent word order when comparing the hypothesis and the reference segments, although its impact will vary depending on whether we assess the fluency or the adequacy of a segment. In the following example we observe that the adjunct of time *on Thursday* occupies an unnatural position in the hypothesis string. In English, adjuncts of time are usually located in front or back position, that is to say, at the beginning of the clause or after the arguments of a verb. Thus, the hypothesis segment shows problems in terms of fluency due to the position of the adjunct; however, in terms of adequacy, we can affirm that the meaning has been conveyed and can be perfectly understood.

Example 42

HYP: *... Putin* ***on Thursday*** *announced that...*

REF: *Putin announced* ***on Thursday****...*

However, in other cases, the position of adjuncts affects fluency in a lower degree or it does not affect fluency at all, as illustrated by Example 43. When dealing with adjuncts of place (*in the state of Alabama south of the United States*) and time (*within 10 days*), their prototypical order is first the adjunct of place, followed by the adjunct of time. However, sometimes their order can be swapped with no negative effects on fluency or adequacy. The hypothesis

segment follows the prototypical order location + time, whereas in the reference string the adjunct of time precedes the adjunct of location. However, neither fluency nor adequacy is affected.

Example 43

HYP: *...15 churches burned down* ***in the state of Alabama south of the United States*** *within 10 days...*

REF: *...Ten churches were burned down in 10 days* ***in the state of Alabama in the southern United States****...*

- Last but not least, we should also deal with unexpected endings at phrase level, which is typical of statistical machine translation, and turn the chunk into an incorrect one. As shown in Example 44, the PP *de la creación de* in the hypothesis sentence is an ungrammatical phrase as it should not finish with the preposition *de*, but it seems to be a consequence of the learning procedure behind the learning of statistical MT.

  Example 44

  SOURCE: *The human act of creating*

  HYP: *El acto humano de la creación* ***de***

  REF: *El acto humano de la creación*

### 4.2.5 Semantic Level

In the previous sections we covered format, lexical items, morphology and syntax and, with no doubt, when analysing those different linguistic issues semantics was not disregarded. However, in the current section we, especially focus on semantics in a more detailed way, both lexical semantic and sentence semantic relations will be analysed.

**Lexical semantic relations.** Lexical semantics becomes very important when using reference translations in order to evaluate MT output. The reason is that when using reference translations we cannot necessarily expect to find exactly the same words in

the hypothesis and the reference segments. On the contrary we must be ready to establish lexical relations which go beyond the word-form (i.e. synonymy, hyperonymy and hyponymy, homonymy and polysemy).

- Synonymy. A relation of synonymy can be established between two words which have almost identical or similar meanings in a particular context, as exemplified by the two synonym verbs *believe* and *think* below.

  Example 45

  HYP: ...*I* ***believe*** *that the situation...*

  REF: ...*I* ***think*** *the situation...*

- Hypernymy and hyponymy. The relation of hyperonymy refers to a word with a general meaning that includes the meanings of other particular words. On the contrary, a relation of hyponymy can be established when one word is included in the meaning of another more general word or is more specific. The example below illustrates these semantic reations: the word *press* is a hyperonym of *papers* and *papers* is a hyponym of *press*.

  Example 46

  HYP: ...*in European* ***papers****...*

  REF: ...*in the European* ***press****...*

- Homonymy. A relation of homonymy is established between two words that have the same spelling and pronunciation but different meanings. Homonyms pose a problem to MT because the context where this word appears is the only thing that can disambiguate the meaning of that word in the sentence. The word *Mass* in the example below illustrates this issue. The source word *Mass* is a homonym in English (if we do not take into account the capital letter) because it can refer both to a religious celebration (as in this case) and to an amount of a substance. The MT engine fails in identifying the correct translation for *Mass* and instead of translating it as *Misa* (religious celebration) it is translated as *Masa* (amount of a substance).

Example 47

SOURCE: *A **Mass** celebrated for the dead*

HYP: *Una **Masa** celebrada por el muerto*

REF: *Una **Misa** celebrada por el muerto*

- Polysemy. A polysemic word is a word that can have different meanings. It differs from homonyms because homonyms have different origins. The problem that polysemy creates to MT is similar to the one created by homonyms, the semantic ambiguity and the possible translations in the target language. This is the case of the word *fault* in English that can refer both to something that is wrong or not perfect and to something wrong that has been done. An example of mistranslation of a word due to polysemy is displayed in the example below, where the source word *fault* is translated in the candidate string as *culpas* instead of *errores*, which would be the correct translation into Spanish.

  Example 48

  SOURCE: *I could understand his English in spite of his grammatical **faults***

  HYP: *podría entender su inglés a pesar de su grammatical **culpas***

  REF: *podría entender su inglés a pesar de sus **errores** gramaticales*

Some other lexical semantic relations could be established such as antonymy, meronymy, etc. but they do not seem to be as frequent and relevant as those shown above when comparing the hypothesis and the referent translation.

Now that lexical semantic relations have been visited, our discussion below is centred on sentence semantics and what prevents a sentence from being partly or fully understood.

**Sentence Semantics.** When comparing a hypothesis and a reference segment in terms of meaning we want to check whether the meaning of both sentences is the same and whether the MT system has been able to capture and express the meaning of the source sentence. However, it turns out that in many cases, and especially when dealing with

very long segments, MT systems fail to express either part or the whole meaning of the sentence. Below we show an excellent example of this issue.

Example 49

HYP: *Rod Larsen said on radio talking Norwegian " that I saw Kai " that " now part of the many expensive comparing Baruod Barrels in May to prevent further ".*

REF: *Roed-Larsen, speaking to the Norwegian NRK radio station said that "the region can now be compared in several respects to a powder keg with a lit fuse. "*

What is the cause of this failure of translation? And even more important, how can we identify this loss of meaning when automatically comparing the hypothesis and reference string? In the data under analysis, basically, the loss of meaning occurs when some parts of the source sentence have not been translated so they are missing in the hypothesis string, there is a mistranslation of a word or a part of the sentence or the word order in the hypothesis is completely wrong. It is undeniable that these problems are also related to syntax; however for the sake of clarity they have been included in a separate section.

Sometimes it is difficult to understand the meaning of a sentence or part of it because there are some parts missing which prevent us from fully understanding it. These problems of comprehensibility range from a slight loss of meaning or weakening in fluency, caused by the lack of some unimportant words or short phrases (i.e. function words such as determiners or punctuation marks), to the total incomprehensibility of the sentence, caused by the lack of important words (i.e. content words such as non-copula verbs, proper nouns, etc.) or whole phrases. In the section aimed at lexical level (see section 4.2.2), we have already covered those cases that involve the lack of a target word, in this section we focus on the lack of whole semantic arguments and/or adjuncts that causes the loss of information when translating from the source to the target language.

The semantic relation that a verb has with its arguments and adjuncts is usually explained by means of Semantic Roles, in other words, Semantic Roles are used to indicate the role that each constituent plays in a sentence. Therefore, when one of those

entities is missing, part of the meaning of the sentence is missing as well. Such is the case in the examples below where *He* (in Example 50) and *Chechens* (in Example 51), playing the Semantic Role of agent (i.e. performer of the action expressed by the verb), are missing in the hypothesis sentences, and as a consequence crucial information in order to understand the meaning of the sentence is lost in the translation process.

Example 50

HYP: ***ø*** *continued "the executive committee discussed...*

REF: ***He*** *continued, "The Executive Committee discussed...*

Example 51

HYP: ....***ø****carrying out an attack in Moscow"...*

REF: ...***Chechens*** *carry out an attack in Moscow*"...

Once the section on semantics is finished, a detailed and clear picture has been drawn of those linguistic phenomena that must be considered, at least in the corpora used, when comparing hypothesis and reference translation.

### 4.3 Findings

The analysis conducted has been of great help not only to confirm those linguistic features that are already taken into account when developing MT evaluation metrics (i.e. synonymy, stemming, word order), but also to highlight other kinds of linguistic information that has not been used so far in automatic MT evaluation metrics. Such information includes the use of hyponymy relations as regards lexical semantics and valency alternations as regards syntax, to mention a couple of examples. This global view of both Spanish and English corpora helped us draw some preliminary conclusions that will be checked in our experiments and the remaining of this work.

- Although some MT metrics try to be language-independent (BLEU, NIST), and therefore disregard linguistic information, our linguistic analysis shows that the use of linguistic information has a key role in order to ensure a wide coverage at both lexical, syntactic and semantic level. Some examples of the importance of

using linguistic information are the possibility of matching abbreviations and acronyms to their full forms; identifying lexical semantic relations such as synonymy, hypernym and hyponymy; or considering different syntactic constructions implying the same meaning, just to name a few.

- There is a need to consider different dimensions of language when assessing MT. Most of the well-known metrics that use linguistic information, such as METEOR, work at lexical level, thus omitting information at syntactic level. However, from a linguistic point of view the combination of linguistic features from different dimensions of language in MT evaluation seems more appropriate given that it would allow dealing with phenomena beyond the lexical level (i.e. phrase, clause and sentence structure).

- The importance of the linguistic features used in MT metrics seems to vary depending on the kind of assessment (i.e. fluency or adequacy). Therefore, when combining linguistic traits, the type of evaluation must be taken into account in order to give more importance to a specific type of linguistic information. Those metrics that try to use a wide range of linguistic features (Giménez and Márquez, 2010b) usually evaluate MT quality in general, thus no difference is established in order to give a higher weight to those linguistic features that are more involved in a specific type of evaluation than others and it is difficult to measure the influence of each metric.

- In addition, as expected and confirmed by the analysis conducted, all languages cannot be assessed using the same parameters, and once again, the type of linguistic traits and their importance may vary depending on the language. For example, linguistic features related to morphosyntax play a more important role when evaluating Spanish than English, because the former shows a richer inflectional morphology than the later.

In order to test whether these conclusions drawn from the linguistic analysis of our corpora do have an influence on the results of MT evaluation, we have developed VERTa, an MT evaluation metric that uses linguistic information at different levels and that is able to use different linguistic features and different parameters depending on the

type of evaluation and language assessed. Next chapter describes VERTa, its modules and its functioning.

# Chapter 5. VERTa: Metric Description

The previous chapter identifies and describes linguistic phenomena that should be addressed when evaluating MT output by means of reference translations. Although covering all those phenomena would be an arduous and difficult-to-approach task, we do consider that the most linguistically relevant ones could be covered using linguistic information and resources. This implies true MT errors and those phenomena that might be wrongly identified as MT errors, but which are not. Thus, the most natural step to confirm our hypothesis was to develop an MT metric, VERTa, to check the influence and suitability of the linguistic information proposed, as well as to evaluate MT output. With this aim in mind, we first propose a classification of the linguistic information required into different application levels (section 5.1) and then the architecture of our MT metric and its different modules (section 5.2) are described. Finally, a brief summary of this chapter is provided in section 5.3.

## 5.1 Organising Linguistic Information

When approaching the design and development of our linguistically-motivated metric, VERTa, we identified several linguistic issues which should be considered when comparing hypothesis and reference segments (see Chapter 4). From these linguistic phenomena, the most relevant ones were selected and classified. Even though most of them are interrelated and interact, they were classified into the following levels for the sake of analysis:

- **Lexical information:** At this level we want to highlight the importance of lexical semantics. Lexical semantics becomes very important when using reference translations in order to evaluate MT output, because we cannot necessarily expect to find exactly the same word-forms in the hypothesis and the references. On the contrary, we must be able to establish more flexible lexical relations that do not only involve the word-form but also semantics such as synonymy (e.g. *believe – think*), hypernymy and hyponymy (e.g. *papers – press*).

- **Morphological information:** Morphology is an important element, especially when dealing with languages with a rich inflectional morphology, such as Spanish, French and Catalan because it helps us to deal with linguistic features such as tense, aspect, mood, number, gender or case. Therefore, by means of morphological features such as tense, we can compare whether the tense used in the hypothesis and the reference translation is the same or varies. On the other hand, both derivational and inflectional morphology are also worth mentioning since they help in identifying words that belong to the same form-based word family thus sharing the same root (e.g. *participate* and *participation)* or lemma (e.g. *helped* and *help*). Moreover, inflectional morphology in combination with syntax (morphosyntax) also plays an important role in the fluency of a sentence. Such is the case of agreement in English, where verb forms in third person singular show agreement with the subject by means of the *–s* ending.

- **Syntactic information:** At this level a couple of issues are considered, the syntactic structure and the word order, both inside the phrase and inside the clause. This level covers those changes that imply a change of grammatical category (e.g. from a NP to a PP), thus a different syntactic structure, and those that do not entail a change in the grammatical category of the units affected but account for the constituent word order. A couple of examples about the syntactic changes mentioned above are the following, where the active-passive alternation is illustrated (Example 52) as well as change in word order inside the phrase (Example 53).

  Example 52

  HYP: *...were assassinated by unknown men...*

  REF: *...unknown men assassinated...*

  Example 53

  HYP: *...a **Moroccan official** source...*

  REF: *... an **official Moroccan** source...*

- **Sentence Semantics information:** This level is centred on sentence semantics and the causes which prevent a sentence from being partly or fully understood. As explained in section 4.2.5, although most of these issues are rooted in syntax, they have been included in a different section for the sake of clarity. Such is the case of the example below where the subject of the sentence realised by the proper noun *Chechens* is missing in the hypothesis sentences, and as a consequence we do not have information on the entity performing the action expressed by the verb.

  Example 54

  HYP: ....ø*carrying out an attack in Moscow"*...

  REF: ...***Chechens*** *carry out an attack in Moscow*"...

In order to combine the above described linguistic information and check its use and influence on MT evaluation, we have decided to develop a similarity metric[31], VERTa.

## 5.2 The Metric Architecture and Description

VERTa is an MT metric that uses reference translations. Thus in order to check the similarity between MT output and the reference translation, VERTa compares each hypothesis segment with the corresponding reference segment according to different types of linguistic information.

VERTa, consists of several modules working at different levels: Lexical Module, Morphological Module, Dependency and Semantic Module. Moreover, we have also added an N-gram Module so as to account for similarity between chunks and a Language Model (LM) Module[32]. In addition, the organisation of linguistic features in different modules or levels allows us to evaluate both adequacy and fluency, thus checking the suitability of linguistic features for both types of evaluation.

In VERTa, each module works first individually and the final score is the Fmean of the weighted combination of the Precision and Recall of each module in order to get the

[31] Note on the terminology used. From now on, to avoid confusion, we will use the following terms: *Metric*, refers to the whole program, VERTa; *Module*, refers to the set of linguistic features per level.
[32] Both Semantic and Language Model Modules are only available for English.

results which best correlate with human judgements (see Figure 11). This way, the different modules can be weighted depending on their importance regarding the type of evaluation (fluency or adequacy) and language evaluated. In addition, the modular design of this metric makes it suitable for all languages. Even those languages that do not have a wide range of NLP tools available could be evaluated, since each module can be used in isolation or in combination. It must be highlighted that the first module applied is the Lexical Module and the matches set by this module are the basis of the alignment. VERTa allows for two possible alignments: the first alignment only takes into account those matches set in the Lexical Module, whereas in the second one the Lexical Module and the Morphological Module work as a team, and as a consequence, the Morphological Module benefits from the matches established in the Lexical one (i.e. word-forms, lemma, synonyms, hypernyms, hyponyms and partial lemma). The type of alignment used will depend on the type of evaluation (see Chapters 6 and 7 for further details).

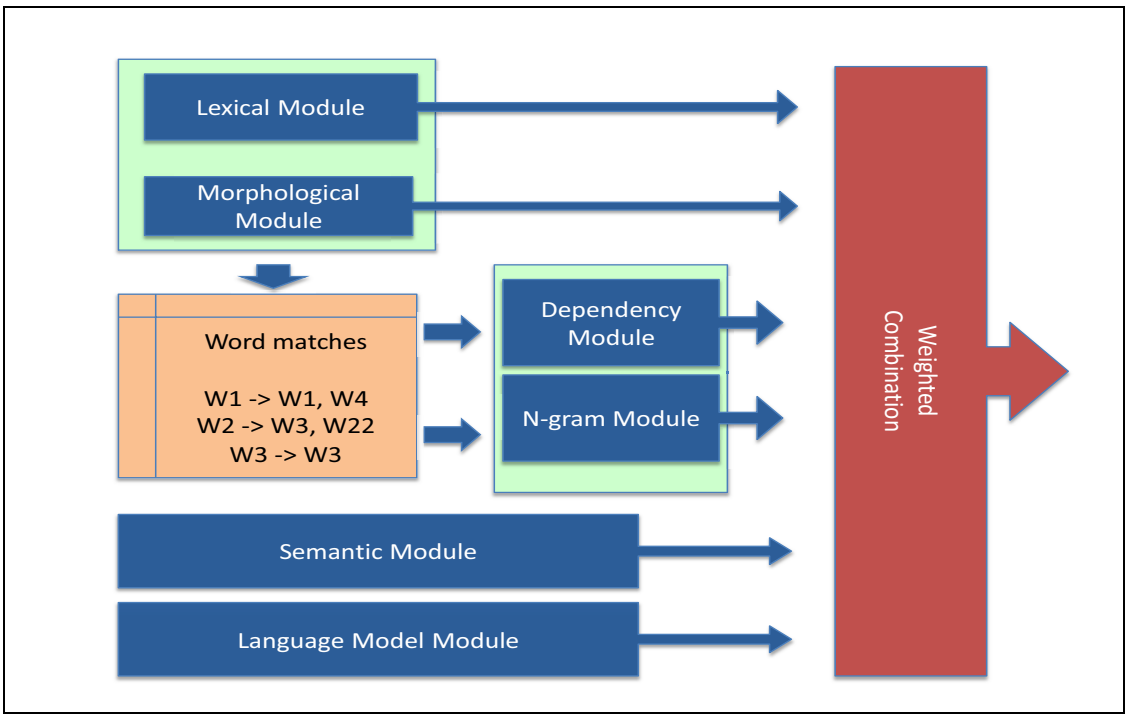



**Figure 11** VERTa's architecture

All modules (except for the Language Model) use a weighted precision and recall over the number of matches of the particular element of each level (words, dependency triples, n-grams, etc.) as shown below.

$$P = \frac{\sum_{\partial \in D} W_\partial * nmatch_\partial(\nabla(h))}{|\nabla(h)|}$$

$$R = \frac{\sum_{\partial \in D} W_\partial * nmatch_\partial(\nabla(r))}{|\nabla(r)|}$$

Where *r* is the reference, *h* is the hypothesis and ∇ is a function that given a segment will return the elements of each level (e.g. words at lexical level and triples at dependency level). *D* is the set of different types of matching. $nmatch_\partial()$ is a function that returns the number of matches of type ∂ (e.g. the number of lexical matches at the lexical level or the number of dependency triples that perfectly match at the dependency level). Finally, *W* is the set of weights [0 1] associated to each of the different types of matching in order to combine the different kinds of matches considered in that level.

VERTa uses the Fmean to combine Precision and Recall measures, if there's more than one reference, the maximum Fmean among all references is returned as the score.

When the scores per module are calculated the final score is a weighted average of the different scores (Fmean) of the modules.

As mentioned before, VERTa works at segment level, comparing the different items of the hypothesis and reference segments from left to right. It must be highlighted that a segment can be composed of one or more sentences. Thus, it could be the case that one segment of the hypothesis contains just one sentence whereas the same segment in the reference has been split into two, as illustrated in Example 55. In order to deal with this fact, first the segment is split into sentences and the linguistic tools used are applied to each sentence, afterwards the metric calculates the score for the whole segment; that is to say, we look for the similarity of all items inside the hypothesis segment as compared to all items in the reference segment.

Example 55

HYP: *He said " two or three days ago on love , throughout the world and , unfortunately , that large number of our youth and women of our happens in the day*

*in what patriotism enemy eyes as gifts and speeches and to exchange reached limit , that we want our enemies that live away from any virtue " .*

REF: *He said , " The international Valentine 's Day passed two or three days ago , and unfortunately , many of our young men and women do on that day what delights our enemy . They exchange gifts and exchange words , and it has reached the point of exchanging kisses . This is what our enemies want for us , that we live far away from virtue . "*

All modules forming VERTa and the linguistic features used are described in detail in the following subsections.

### 5.2.1 Lexical Module

The Lexical Module compares lexical items from the hypothesis segment with those in the reference segment. The approach followed in this module is inspired by METEOR (Banerjee and Lavie 2005) in the sense that the module relies on lexical items and lexical semantic relations. However, while the recent versions of METEOR (Denkowsi and Lavie 2011/2014) deal with semantics by means of synonymy and paraphrase tables, our metric does not only use synonymy[33] but it also makes good use of other lexical semantic relations such as hypernymy and hyponymy avoiding the use of paraphrase tables which have to be built up for each language and domain. Moreover, VERTa also employs the information provided by lemmas and partial lemmas, whereas METEOR relies only on stemming. In addition, we also apply a system of weights on the different matches established depending on their importance as regards semantics.

Table 5 provides the list of features used in the Lexical Module. The matching process follows.

[33] For further details on the resources used to obtain information regarding synonyms, hypernyms, hyponyms and lemmas, please refer to section 3.2.2.

| | Match | Examples | |
|---|---|---|---|
| | | **Hypothesis** | **Reference** |
| 1 | Word-form | *east* | *east* |
| 2 | Lemma | *is_BE* | *are_BE* |
| 3 | Synonym | *believed* | *considered* |
| 4 | Hypernym | *barrel* | *keg* |
| 5 | Hyponym | *keg* | *barrel* |
| 6 | Partial lemma | *danger* | *dangerous* |

**Table 5** Lexical matches and examples

First, VERTa looks for word-form matches between lexical items in the hypothesis and reference segments. With those words that cannot be matched, VERTa tries to establish matches using lemmas synonyms. Then, with the words left the metric looks for similarities between synonyms. If there are still unmatched words, VERTa tries similarities between hypernyms/hyponyms and finally partial lemmas (it checks whether the first 4 letters of one lemma in the hypothesis segment can match with the first 4 letters in the reference segment). The use of the linguistic features described above may vary depending on the type of evaluation and language evaluated. In addition, each type of match can receive a specific weight according to their relevance in the type of evaluation performed. Please refer to Chapter 6, 7 and 9 for further details.

### 5.2.2 Morphological Module

The Morphological Module allows for combining lexical and morphological information or using morphological information by itself. When used in the combinatory fashion, the module is based on the matches established in the Lexical Module (see section 5.2.1) in combination with PoS tags from the annotated corpus[34](see section 3.2.2 for further details).

On the other hand, when only morphology information is used, it is only based on PoS matches between the hypothesis and reference segments. The aim of this module is to

[34] The English corpus has been annotated by the Standford Log-Linear Part of Speech Tagger (Toutanova et al. 2003), included in the Standford CoreNLP suite, and the Spanish corpus by Freeling (Padró and Stanilovsky 2012).

compensate for the broader coverage of the Lexical Module, preventing matches such as *invites* and *invite*, which although similar in terms of meaning, differ on their morphological information. Therefore, this module seems to be more appropriate to assess the fluency of a segment rather than its adequacy. In addition, although this module may not play a key role when assessing English output, it might be particularly useful when evaluating languages with a richer inflectional morphology (e.g. Romance languages).

In line with the Lexical Module, the Morphological Module establishes matches between items in the hypothesis and the reference sentence and different weights can also be assigned to each type of match (see Chapters 6, 7 and 9). However, instead of comparing single lexical items as in the previous module, in its combinatory fashion, this module compares pairs of features in the order established in Table 6.

| | **Match** | **Examples** | |
|---|---|---|---|
| | | **Hypothesis** | **Reference** |
| 1 | (Word-form, PoS) | (he, PRP) | (he, PRP) |
| 2 | (Synonym, PoS) | (VIEW, NNS) | (OPINON, NNS) |
| 3 | (Hypern., PoS) | (PUBLICATION, NN) | (MAGAZINE, NN) |
| 4 | (Hypon., PoS) | (MAGAZINE, NN) | (PUBLICATION, NN) |
| 5 | (Lemma, PoS) | can_(CAN, MD) | could_(CAN, MD) |

**Table 6** Morphological Module matches

Therefore, first, the metric compares the word-form and PoS of one lexical item in the hypothesis sentence with the corresponding values of another one in the reference sentence. With the pending words, VERTa checks whether two lemmas appear as synonyms in WordNet and also compares the PoS annotation of each word-form. Then the metric examines if one lemma is the immediate hyperonym/hyponym of the other in WordNet and also compares the PoS tags. Finally, the metric compares the lemma and PoS of a specific word in the hypothesis sentence to the corresponding values of another word in the reference sentence, as well as their PoS. Although this last type of match might not be very useful in English (i.e. in our corpus the only instance found is the example shown in Table 6), in Spanish it might help to avoid misleading matches such as verb forms *era (was - imperfect)* and *fue* (*was - preterite*).

In addition, this module allows for a last type of match that only uses PoS information, disregarding the use of matches established in the Lexical Module.

### 5.2.3 Dependency Module

Once both Lexical and Morphological sections which worked at lexical level have been covered, we now move to the Dependency Module which accounts for the phrase and clause structure. By means of this module the metric is able to capture relations between sentence constituents regardless of their position inside the sentence (see Example 56), as well as similarities between semantically comparable expressions that show a different syntactic structure.

Example 56

HYP: *Ramallah ( West Bank ) 2-15 ( AFP ) - The executive committee of the PLO said today Wednesday that…*

REF: *Ramallah ( West Bank ) 2/15 (AFP) -- Today, Wednesday, the Executive Committee of the Palestine Liberation Organization expressed the opinion...*

In Example 56, the adjunct of time *today Wednesday* occupies different positions in the hypothesis and reference strings. In the hypothesis it is located after the verb, whereas in the reference, it is placed at the beginning of the sentence, preceding the subject *The executive committee of the PLO*. By means of the dependency analysis, we can state that although located differently inside the sentence, both subject and adjunct depend on the verb (see Table 7).

| Hypothesis | Reference |
|---|---|
| nsubj(committee, said) | nsubj(Committee, expressed) |
| tmod(today, said) | tmod(Today, expressed) |

**Table 7** Comparison between hypothesis and reference triplets

Therefore, the use of dependencies proves to be effective in order to establish similarities between equivalent sentences which contain the same constituents but in different positions.

This module works at sentence level and follows the approach used by Owczarzack et al. (2007a/b) and He et al. (2010) with some changes and linguistic additions in order to adapt it to our metric combination. One of the differences between the above mentioned proposals and ours is that they used an LFG parser and MALT parser respectively, whereas the parser used in VERTa is the Stanford parser (de Marneffe et al. 2006) for English and Freeling (Lloberes et al. 2010) for Spanish (please refer to section 3.2.2 for further details). The reason why this Standford parser is used is because after conducting an evaluation (Comelles et al. 2010) where the performance of several dependency parsers was assessed (Stanford, DeSR (Attardi 2006), MALT (Nivre 2006), Minipar (Linn 1998) and RASP (Briscoe et al. 2006)) this proved to be the best in terms of linguistic quality. As regards Freeling, we opted for this tool because it is a knowledge-based parser and does not require any kind of training, which made it more suitable for the type of data used in Spanish (see section 3.2.1.5 for further details).

Similar to the Morphological Module, the Dependency Module also relies first on those matches established at lexical level − word-form, synonymy, hypernymy, hyponymy, lemma and partial lemma − in order to capture lexical variation across dependencies and avoid relying only on surface word-form.

Then, by means of flat triplets with the form Label(Head, Mod) obtained from the parser, four different types of dependency matches have been designed (see Table 8) and weights can also be assigned to each type of match (see Chapter 6, section 6.4.2, and Chapter 7, section 7.4.1 for weights proposed).

| | Match Type | Match Description |
|---|---|---|
| 1 | Exact | Label1=Label2<br>Head1=Head2<br>Mod1=Mod2 |
| 2 | No_label | Label1≠Label2<br>Head1=Head2<br>Mod1=Mod2 |
| 3 | No_mod | Label1=Label2<br>Head1=Head2<br>Mod1≠Mod2 |
| 4 | No_head | Label1=Label2<br>Head1≠Head2<br>Mod1=Mod2 |

**Table 8** Dependency matches

These matches are applied in the order established in Table 8. First VERTa looks for Exact matches (i.e. triples in the hypothesis and reference segments are identical). Then, the metric moves to the No_label match, thus comparing triples that show identical head and modifier but different label, as shown in below.

Example 57

HYP: *...**all** Palestinian political **parties**...*

REF: *...**all** the Palestinian political **parties**...*

predet(**parties**,**all**)=det(**parties**,**all**)

With the triples left, VERTa tries to establish matches between those triples that show the same label and head but different modifier, as illustrated next.

Example 58

HYP: *...the situation more **difficult** and complicated and **serious**...*

REF: *...the situation is more **difficult**, complicated and **dangerous**...*

**conj_and**(**difficult**,dangerous) = **conj_and**(**difficult**,serious)

Finally, the metric looks for triples that share the same label and modifier but different head:

Example 59

HYP: ...*He **said** "I **believe** that the situation..."*

REF: ...*He **added** "I **think** ø the situation..."*

**ccomp**(said,**believe**)=**ccomp**(added,**think**)

This last type of match, the No_head match, was also proposed by Owczarzak et al. (2007a/b); however, He et al. (2010) disregarded this type of match in their proposal. Although no arguments were given for such a decision, we might think that it did not correlate well with human judgements. In our metric, we decided to use this fourth type of match because we were interested in checking its suitability, not only as regards correlation with human judgements but also regarding linguistic analysis (see experiments in Chapter 6, section 6.4.1.3).

In addition, and following He et al. (2010)'s approach, dependency labels are given different weights depending on their suitability and importance depending on the type of evaluation (see Chapter 6, section 6.4.3, and Chapter 7, section 7.4.2).

As regards the way the final score for this module is calculated, a couple of parameters are considered: the type-of-match weight and the dependency-relation weight. Therefore, each triple match combines the weight given to the type of match and the weight assigned to the dependency label. Then matches are added up and precision and recall are calculated.

Finally, a set of language-dependent rules has been added with two goals: 1) capturing similarities between different syntactic structures conveying the same meaning, in case the dependency matches did not capture them; and 2) restricting certain dependency relations (e.g. subject word order when translating from Arabic to English). Although these rules are implemented in the Dependency Module, they also affect sentence

semantics. In order to cover these two goals, the set of language-dependent rules is further divided into two types of rules.

The first type of rules is applied at phrase and clause level. The former affects modifiers inside the noun phrase and the latter valency alternations regarding the verb and its complements. As for the structure inside the noun phrase, we cover a couple of phenomena:

a) The similarity between an adjective or noun premodifiying a noun and an of-prepositional phrase postmodifying it, as exemplified below (see Example 60).

   Example 60

   HYP: *...between the* ***ministries of interior****...*

   REF: *...between the two* ***interior ministries****...*

   HYP_prep_of(ministries,interior) = REF_nn (ministries,interior)

   Although their labels differ, this couple of triples must be considered as an Exact match due to their semantic similarity. Otherwise we would penalise a couple of structures which are equal from a semantic point of view. By means of the rules *(Lprep_of-Lamod,X,X): 1.0* and *(Lamod-Lprep_of,X,X): 1.0*, or *(Lprep_of-Lnn,X,X): 1.0* and *(Lnn-Lprep_of,X,X):1.0*, the total similarity between these two structures is granted. This rule states that when a triple in the hypothesis segment and a triple in the reference segment share both head (*X*) and modifier (*X*), but their dependency labels (*L*) are *prep_of* and *amod* or *prep_of* and *nn* these triples must be considered as an Exact match.

b) The similarity of a possessive structure expressed by means of a possessive *'s* and a possessive structure expressed by means of an of-prepositional phrase, as exemplified in the following example:

   Example 61

   HYP: ... ***Mark's mother****...*

   REF: *...the* ***mother of Mark****...*

HYP_poss(mother,Mark) = REF_prep_of(mother,Mark)

The rule *(Lposs-Lprep_of,X,X): 1.0* is responsible for covering this similarity and assigning the maximum weight, since the meaning they express is identical. This rule states that when a triple in the hypothesis segment and a triple in the reference segment share both head (*X*) and modifier (*X*), but their dependency labels (*L*) are *poss* and *prep_of*, these triples must be considered as an Exact match.

At a clause level, structures involving verbal complements must be under consideration:

a) The first of these structures is the active-passive alternation. As shown in the example below, although syntactically different, both structures share the same meaning.

Example 62

HYP: *After meeting* ***the Moroccan news agency published*** *a joint statement...*

REF: *A joint statement was* ***published*** *(...)* ***by the Moroccan news agency****...*

HYP_nsubj(published,agency) = REF_agent(published,agency)

Similar to the pair of dependencies dealing with modifiers, *nsubj-agent* and *nsubjpass-dobj* labels must be considered identical and thus, the previous pairs of triples must be scored as an Exact match. The rules *(Lnsubj-Lagent,X,X): 1.0* and *(Lnsubjpass-Ldobj,X,X)*: 1.0, respectively, cover these similarities. The first rule states that when a triple in the hypothesis segment and a triple in the reference segment share both head (*X*) and modifier (*X*), but their dependency labels (*L*) are *agent* and *subject*, these triples must be considered as an Exact match. The second rules states that when a triple in the hypothesis segment and a triple in the reference segment share both head (*X*) and modifier (*X*), but their dependency labels (*L*) are *nsubjpass* and *dobj*, these triples must be considered as an Exact match

b) In addition, the dative-ditransitive alternation is also considered, taking into account both recipient and beneficiary, as shown in Example 63 and Example 64, respectively.

Example 63

HYP: *He **gave Mark** a book.*

REF: *He **gave** a book **to Mark.***

HYP_iobj(gave,Mark) = REF_prep_to(gave,Mark)

Example 64

HYP: *He **bought Mary** a book.*

RE: *He **bought** a book **for Mary.***

HYP_iobj(bought,Mary) = REF_prep_for(bought,Mary)

Rules *(Liobj-Lprep_to,X,X): 1.0* and *(Liobj-Lprep_for,X,X): 1.0* will account for such similar structures. The former states that when a triple in the hypothesis segment and a triple in the reference segment share both head (*X*) and modifier (*X*), but their dependency labels (*L*) are *iobj* and *prep_to*, these triples must be considered as an Exact match; whereas the latter states that when a triple in the hypothesis segment and a triple in the reference segment share both head (*X*) and modifier (*X*), but their dependency labels (*L*) are *iobj* and *prep_for*, these triples must be considered as an Exact match.

The second type of rules, those more restrictive, allow for controlling and restricting the most flexible type of matches (No_label, No_mod and No_head). As explained in Chapter 4, the reordering of some constituent might still be a problem for some MT engines. This turns into a critical issue when failing in reordering affects immediate constituents such as the subject or the object, since this may lead to a completely wrong translation. Furthermore, these rules can also help to control wrong translations of these key dependency relations (see Example 65).

Example 65

HYP: ***An*** *said that ".the poor manner..."*

REF: ***Jazairi*** *said that "the way..."*

HYP_nsubj(said,**An**) = REF_nsubj(said,**Jazairi**)

The No_mod match would allow for the similarity between these two dependency triples, which is obviously wrong and affects a crucial dependency relation. In order to avoid this issue, a set of restrictive rules on the subject have been developed and applied:

L(nsubj,X,X): 1.0

L(nsubj,X,O): 0

L(nsubj,O,X): 0

The rules state that a triple showing the label (L) *nsubj* will only match another triple with the same label if the head and the modifier coincide (*X,X*), otherwise (*X,O* and *O,X*) the possible match will be disregarded.

### 5.2.4 N-gram Module

The N-gram Module matches chunks in the hypothesis and reference segments and can rely either on a) the matches set by the Lexical Module, which allows us to work not only with word-forms (as BLEU does) but also with synonyms, lemmas, partial lemmas, hypernyms and hyponyms as shown in Example 66, where the chunks [*the situation in the area*] and [*the situation in the region*] do match, even though *area* and *region* do not share the same word-form but a relation of synonymy; or b) the matches set by the Morphological Module, in other words combining lexical matches and PoS information or using PoS information isolated.

Example 66

HYP: ... *the situation in the* ***area****...*

REF: ... *the situation in the* ***region****...*

Chunks length may go from bigrams to sentence length, depending on the type of evaluation (see sections 6.5 and 7.5). The use of this module allows the combination of both linguistic and statistical approaches and enables us to deal with word order inside the sentence by means of a more simple approach than the parsing of constituents.

### 5.2.5 Semantic Module

Semantics plays an important role in the evaluation of adequacy. This has also been claimed by (Lo and Wu 2010) who report that their metric based on Semantic Roles (SR) outperforms other well-known metrics when adequacy is assessed. The Semantic Module in VERTa does not use information on SRs since dependency relations are thought to be halfway between syntax and semantics, thus one of our hypotheses is that the Dependency Module could also provide information in this sense. However, the Semantic Module uses other semantic information at both lexical and sentence level: NEs, Time Expressions and Sentiment analysis.

Regarding NEs, we use Named Entity recognition (NER) and Named Entity linking (NEL). Following previous NE-based metrics (Reeder et al. 2011 and Giménez 2008a) the NER component captures similarities between NEs in the hypothesis and reference segments. In order to identify NEs we use the Supersense Tagger (Ciaramita and Altun 2006) for English. On the other hand the NEL component focuses only on those NEs that appear on Wikipedia, which allows for linking NEs in the hypothesis and reference segments regardless of their external form. Thus, *EU* and *European Union* will be captured as the same NE, since both of them are considered as the same organisation in Wikipedia. The NEL component uses a graph-based NEL tool inspired by Hachey et al. (2011) which links NEs in a text with those in Wikipedia pages.

As regards the Time Expressions (TIMEX) component, it matches temporal expressions in the hypothesis and reference segments regardless of their form. The tool used is the Stanford Temporal Tagger (Chang and Manning 2012) which recognizes not only points in time but also duration. By means of the TIMEX component, different syntactic structures conveying the same time expression can be matched, such as *on February 3rd* and *on the third of February*.

Finally, Sentiment analysis has been added using the dictionary strategy described in Atserias et al. (2012). Sentiment analysis provides information regarding the contextual polarity of the sentence, whether it has a positive or negative connotation.

### 5.2.6 Language Model Module

The Language Model (LM) Module works differently from the rest of modules, in the sense that it neither tries to find similarity matches between the hypothesis and reference segments, nor tries to compare them. This module is only applied to the hypothesis segment and uses a language model to calculate the degree (log probability) to which the hypothesis segment is expected compared to what occurs in the corpus used to build the language model. A language model assigns a probability to a sequence of words (N-grams), thus it is possible to obtain the most frequent N-grams for a specific domain. By using a language model we aim at accounting for those segments that, even being syntactically different from their corresponding reference translations, are still fluent (see Example 35 in Chapter 4); in other words, we will be able to check the correct construction and plausibility of the hypothesis, even if it is very different or not included in any of the reference segments.

## 5.3 Summing Up

This chapter has presented the MT metric developed in the current thesis so as to check the suitability of different linguistic features, separately and in combination, with regard to MT output evaluation. The classification of linguistic features used has been reported as well as the architecture of the MT metric developed, VERTa. This MT metric contains different modules: Lexical, Morphological, Dependency, N-gram, Semantic and Language Model Modules. These modules and the linguistic features used in each of them have been detailed, together with a description of the way they interact and showing that they can be combined depending on the type of evaluation performed.

The next chapter describes the experiments performed with VERTa in order to test all linguistic features included, the performance of the different modules, both individually and in combination, to evaluate adequacy.

# Chapter 6. Experiments on Adequacy

This chapter describes the experiments conducted and results obtained in order to study and test the suitability of the linguistic features used in VERTa, the influence of each module (see section 5.2 for further details on each module) and the best way to combine them in order to evaluate adequacy; that is to say, the degree to which the information present in the input sentence is also communicated in the output sentence. Following the aim of this thesis, experiments conducted in this chapter take correlation coefficients as a point of departure and focus on providing linguistic evidence, supported with examples, of the suitability of those linguistic features used and the influence of each module and their combination. Thus, so as to perform such a fine-grained evaluation of the linguistic information, experiments are carried out at segment level.

This chapter has been organised as follows: section 6.1 describes the data used in the experiments conducted, section 6.2 analyses the features used in the Lexical Module; section 6.3 explores the use of the Morphological Module and its features; section 6.4 examines the Dependency Module; section 6.5 explores the N-gram Module; section 6.6 deals with the module aimed at semantics; section 6.7 discusses the best combination of modules used to evaluate adequacy; finally, findings are reported in section 6.8[35].

## 6.1 Data

So as to perform these experiments we used part of the development data provided in the MetricsMaTr 2010 shared-task[36] (see section 3.2.1.1 for further details). From the data provided by the organization we used 100 segments (Arabic to English) of the NIST Open-MT06 data, the MT output from 8 different MT systems (a total of 28,000 words approximately) and 4 reference translations. The human judgments used were based on adequacy (7-point scale, straight average). In order to calculate correlations at segment level Pearson correlation was applied between our metric and the adequacy judgments. All segments were taken into account regardless of the system providing

[35] The LM Module, presented in Chapter 5, will be used in the Experiments on fluency since it is fluency oriented.

[36] http://www.nist.gov/itl/iad/mig/metricsmatr10.cfm.

them and the evaluation was performed at segment level in order to conduct a more detailed study.

## 6.2 Lexical Module

Similar to most of the lexical metrics used nowadays such as METEOR (Denkowsi and Lavie 2014), SIA (Liu and Gildea 2006), M-BLEU (Agarwal and Lavie 2008), TERp (Snover et al. 2009) and ATEC (Wong and Kit 2008/2010), VERTa also uses information at lexical level. This information covers similarity between word-forms, lemmas, synonyms, partial lemmas and hypernyms/hyponyms. The metric seeks for matches between the hypothesis and reference segments taking into accout this linguistic information.

Section 6.2.1 describes the traditional types of matches used and, in the same line as one of the latest versions of METEOR, the different types of linguistic knowledge used are not given the same importance in terms of weight. Therefore, a couple of questions arose:

- Would other linguistic features improve the performance of our metric?
- Should all linguistic features receive the same importance in terms of weights?

Section 6.2.2 explores the use of hypernyms and hyponyms, as a new linguistic feature to improve the performance of our metric. Section 6.2.3 analyses the weights assigned to each linguistic feature according to their importance in the evaluation of adequacy. Finally, section 6.2.4 presents a summary of the Lexical Module and the linguistic features used.

### 6.2.1 Traditional Types of Matches

The word-form is obviously the most basic unit of comparison between lexical items, when comparing hypothesis and reference segments. However, according to most of the lexical similarity metrics, there are other lexical relations that must be taken into account, such as similarity between lemmas, synonyms and partial lemmas or stems. Some of the examples below show such importance. Example 67 shows an example of lemma match: although *pressure* and *pressures* show different number, they do share

the same lemma, allowing therefore for the transmission of meaning. Besides, Example 68 illustrates the lexical semantic relation of synonymy that can be established between *believed* and *considered*, and *area* and *region*, respectively. Finally, Example 69 illustrates a partial lemma match, since the words *invited* and *invitation* share the same root, thus part of their lemma.

Example 67

HYP: *...we talked about international* ***pressure*** *and threats to cut off aid to the Palestinian people".*

REF: *...we discussed the international* ***pressures*** *and threats to cut aid to the Palestinian people."*

Example 68

HYP: *Terje Rod Larsen former UN envoy to Middle East* ***believed*** *that the situation in the* ***area****...*

REF: *Terje Roed-Larsen, the former United Nations Middle East envoy,* ***considered*** *the situation in the* ***region****...*

Example 69

HYP: *... Putin* ***invited*** *Hamas...*

REF: *... Putin's* ***Invitation*** *to Hamas*

From a linguistic point of view, the use of such features seemed to be relevant; however, their impact on the whole corpus could only be checked by assessing whether they helped to improve the correlation with human judgements. To this aim, correlations were calculated taking into account a 4-reference scenario (see Table 9) and a single-reference scenario (see Table 10). In order to test their influence and their impact in terms of correlation with human judgements, the feature word-form was taken as the starting point and it was combined with the rest of linguistic features (i.e., lemma, synonymy and partial lemma). In this sense, linguistic features were given the same importance by means of assigning the same weight to each of them.

| Linguistic Features | Pearson Correlation |
|---|---|
| Word_Only | 0.5766 |
| Word + Lemma | 0.7212 |
| Word + Lemma +Synonymy | 0.7399 |
| Word + Lemma + Synonymy + Partial lemma | **0.7418** |

**Table 9** Influence of linguistic features in a 4-reference scenario

| Linguistic Features | Pearson Correlation | | | |
|---|---|---|---|---|
| | **Ref. 1** | **Ref. 2** | **Ref. 3** | **Ref. 4** |
| Word_Only | 0.5191 | 0.5085 | 0.5339 | 0.5240 |
| Word + Lemma | 0.6470 | 0.6471 | 0.6487 | 0.5967 |
| Word + Lemma +Synonymy | 0.6891 | 0.6683 | 0.6810 | 0.6226 |
| Word + Lemma + Synonymy + Partial lemma | **0.6957** | **0.6689** | **0.6798** | **0.6344** |

**Table 10** Influence of linguistic features in a single-reference scenario

As shown in Table 9 and Table 10, the more linguistic features used, the higher the correlation with human judgements. The correlations obtained, indicate that the most valuable addition is that of information regarding lemmas and synonyms, since a remarkable improvement in correlation is achieved (0.1446 and 0.0187, respectively). The addition of information regarding partial lemmas also increases the correlation but does not show such a strong influence.

As stated in the introduction of this chapter, most of the current metrics use information at lexical level. Most of them use information regarding synonyms and stemming (METEOR 1.3 (Denkowsi and Lavie 2011), SIA (Liu and Gildea 2006), M-BLEU (Agarwal and Lavie 2008), TERp (Snover et al. 2009) ATEC (Wong and Kit 2008/2010) (see Chapter 2 for details). However, as observed from the study performed in our corpus (see Chapter 4), there are other lexical semantic relations that should be considered, i.e. hyperonyms and hyponyms. The study of this new type of linguistic information was of our interest, mainly because it is not used by any other MT metric and instances of such relations were found in our corpus, which seemed to be an indicator of their relevance.

### 6.2.2 Use of Hyponyms and Hypernyms

On the light of analysing the impact of using different linguistic knowledge in our metric we decided to use other kinds of semantic relations. After analysing the corpus of development we decided to test the semantic relations of hyponymy and hypernymy, because we found several instances of them (see section 4.2.5). However, we are also aware that using too much information may actually be confusing for the metric and then work in its detriment. Therefore, we wanted to check whether these new types of linguistic information could help to improve the performance of the metric or, on the contrary, they could cause more noise and therefore, decrease its performance.

Several levels of hypernymy and hyponymy were tested in order to decide which correlated best with human judgements:

- Multilevel: All possible levels of hypernyms/hyponyms are taken into account.
- Direct: A relation of direct hypernymy/hyponymy can be established regardless of the word sense.
- MFS Direct: The relation of direct hypernymy/hyponymy is restricted to that of the most frequent word sense (MFS). Since no word-sense disambiguation has been used, the first hypernym/hyponym found is considered to be the most frequent one, as stated in http://wordnet.princeton.edu/wordnet/man/wndb.5WN.html.

| Hypernym/Hyponym Relation | Pearson Correlation |
|---|---|
| Multilevel | 0.7376 |
| Direct | 0.7386 |
| MFS Direct | **0.7415** |

**Table 11** Use of hypernyms and hyponyms in a 4-reference scenario

A thorough analysis of the scores obtained (see Table 11) showed that restricting hypernyms and hyponyms to the most frequent word sense helped in reducing noise and as a result results obtained improved (from 0.7376 to 0.7418). On the contrary, disregarding any kind of restriction or disambiguation made the metric match certain words which, although being hypernyms and/or hyponyms, did not show such a semantic relation in the domain analysed. As expected and confirmed by our

experiments (see Table 11), those relations of hypernymy/hyponymy which correlated best with human judgements were those that used a direct hypernym/hyponym of the most frequent word sense. The most natural next step, therefore, was adding this new linguistic feature to the whole metric in order to check its performance. However, as shown in Table 12, the use of hypernyms and hyponyms does not seem to help in improving the metric, on the contrary, their use implies a slight drawback to its performance (-0.0003).

| Linguistic Features | Pearson Correlation |
|---|---|
| Word + Lemma + Synonymy + Partial lemma + Hypernymy & Hyponymy | 0.7415 |
| Word + Lemma + Synonymy + Partial lemma | **0.7418** |

**Table 12** VERTa with and without hypernyms and hyponyms in a 4-reference scenario

The analysis of the data conducted evidenced that on the whole the use of hypernyms did not mean a better correlation with human judgements. However, in some cases the fact of using hypernyms and hyponyms had a positive effect, as shown in Example 70, where the words *now* and *present* turn into a positive match. Without the use of hypernyms such a relation would be disregarded.

Example 70

HYP: ... *investigators can **now** only speculate about the motives of...*

REF: ... *at **present**, investigators can only guess the motives of...*

This fact, as well as the short distance in terms of correlation between the version of the metric using hypernynms/hyperonyms and the one without them, made us think that, although such a semantic relation might not be helpful when more than one reference was used, it might be the case that it was helpful when only one reference was available. Therefore, we set experiments using just one reference translation.

| Linguistic Features | Pearson Correlation | | | |
|---|---|---|---|---|
| | **Ref. 1** | **Ref. 2** | **Ref. 3** | **Ref. 4** |
| VERTa NO Hyps/Hypos | 0.6957 | **0.6689** | **0.6810** | 0.6344 |
| VERTa with Hyps/Hypos | **0.6985** | 0.6660 | 0.6808 | **0.6391** |

**Table 13** VERTa with and without hypernyms and hyponyms in a 1-reference scenario

As shown in Table 13, the use of hypernyms and hyponyms in a 1-reference scenario leads to a slightly positive correlation between the scores obtained by VERTa and human judgements in two of the four references available. It seems therefore, that the use of hypernyms and hyponyms should not be entirely disregarded when only one reference is available, as it might provide a broader coverage at lexical level and it might depend on the data used, which cannot be anticipated before applying VERTa. In addition, and considering those results obtained when MFS hypernymy/hyponymy was used, if a Word Sense Disambiguation (WSD) system was applied, this might also result in an improvement of the metric.

### 6.2.3 Use of Weights

METEOR 1.3 opts for assigning different weights depending on the linguistic information used in order to match lexical items. The results reported by Denkowski and Lavie (2011) are based on the correlations with human judgements. From a linguistic point of view, it seems also quite appropriate to provide different weights to the lexical matches depending on the kind of information used.

A synonym is, according to the Collins English Dictionary, “a word or that means the same or nearly the same as another word”. Therefore, it seems quite clear that similarity between word-forms and similarity between synonyms should be considered equal and assigned the same weight; however, some doubts arose as for similarity between lemmas, partial lemmas and hypernyms/hyponyms. In order to check their importance, different weights were applied resulting in those shown in Table 14. Although results show that the impact of using different weights depending on the type of match is not crucial, they do help to improve the correlation of VERTa with human judgements, as stated in Table 15, where correlations with the same weight and different weights are compared.

| Linguistic Features | Weight |
|---|---|
| Word-form | 1 |
| Synonymy | 1 |
| Hypernym | 1 |
| Hyponym | 1 |
| Lemma | 0.8 |
| Partial Lemma | 0.6 |

**Table 14** Weights assigned to each linguistic feature

| VERTa's Performance | Pearson Correlation |
|---|---|
| Equal weights | 0.7418 |
| Different weights | 0.7438 |

**Table 15** Weights comparison in a 4-reference scenario

As results show, correlation with human judgements improves feebly when linguistic features are assigned different weights, although it does not imply a big difference. From a linguistic point of view, it seemed quite arguable that full credit should be assigned to lemma and, up to a point, to partial lemma features. After analysing data, it has been found that these types of matches might also appear in structures that have not been correctly translated and which slightly affect the meaning of the sentence, as shown in Example 71.

Example 71

HYP: *...Moslems that they may live with their religious".*

REF: *"...Muslims must live with their religion."*

In Example 71, the use of the partial lemma relates the forms *religious* and *religion* which, although sharing the same root, appear in two different syntactic structures, one of which is difficult to understand. Unfortunately, this affects the meaning of the sentence negatively. On the other hand, some instances have also been found that indicate the opposite. It seems, therefore, that a larger amount of data is required to reach a final decision on the final weights for lemma and partial lemma features.

It should also be highlighted that according to correlations with human judgements, when hypernyms and hyponyms are used they should be assigned a higher weight than that assigned to lemmas, indicating then that the matching between two semantically-related words is preferred to the matching between two words sharing the same lemma but showing a different inflectional morphology or lexical category.

### 6.2.4 Summing Up

This section has dealt with the use of linguistic features at lexical level to evaluate the adequacy of a hypothesis segment. The Lexical Module contains information regarding the word-form, lemma, synonymy, hypernymy, hyponymy and partial lemma. These linguistic features have been assigned different weights depending on their importance in terms of meaning; thus, word-form and synonyms receive the maximum weight (1), whereas lemma and partial lemma are assigned lower weights, 0.8 and 0.6 respectively. All these weights have been assigned according to correlations with human judgements; however, more data would be needed in order to reach final weights.

Apart from the traditional linguistic features, the use of hypernyms and hyponyms has also been studied. Although these relations did not show a good correlation in a 4-reference scenario, they should not be entirely disregarded when just 1 reference is available, and in that case, they should be assigned the maximum weight.

The following section is aimed at discussing the linguistic features used in the Morphological Module.

## 6.3 Morphological Module

During the last decade, several metrics have used PoS and morphosyntactic information. In order to assess the fluency of a segment, Hamon and Rajman (2006) developed the X-score, a metric based on morphosyntactic features to assess fluency. In addition, Giménez (2008a) developed the SP metric which calculated overlapping over a particular type of PoS or over all PoS types in order to assess translation quality. More recently, Fisher et al. (2012) proposed TerrorCat, an MT evaluation metric aimed at quantifying translation quality based on the frequencies of different error categories, one of them being the number of errors according to the PoS.

PoS tags provide information about the lexical category of the word, as well as its morphosyntactic features. These features connect morphology and syntax, thus playing a key role in the grammaticality of the sentence. Our hypothesis, therefore, is that this second module of VERTa should be more fluency oriented, at least when dealing with English, and should restrict the Lexical Module, which allows for a broader coverage and is crucial when assessing segments in terms of adequacy. However, experiments were also carried out to check the module's performance when assessing adequacy and find linguistic evidence that supported our initial hypothesis.

This section is organised as follows: 6.3.1 describes the similarity matches used in this module and provides linguistic evidence to justify their use; section 6.3.2 compares results obtained by this module when evaluating adequacy with those obtained by the Lexical Module; finally, section 6.3.3 summarizes the use of the Morphological Module and its linguistic features.

### 6.3.1 Similarity Matches

The Morphological Module combines lexical and morphological information. It is based on some of the matches set in the Lexical Similarity Module, namely word-form, synonyms, hypernyms and hyponyms (whenever necessary) in combination with PoS tags from the annotated corpus (see section 3.2.2.1 for details on the tools used). In fact, instead of comparing single lexical items as in the Lexical Module, this module works with pairs of features compared as established in Table 16. In addition, this module also allows for matching PoS tags on their own. As for the weights assigned to each type of match, they follow the same parameters as the Lexical Module.

| **Match** | **Examples** | |
|---|---|---|
| | **Hypothesis** | **Reference** |
| (Word-form, PoS) | (he, PRP) | (he, PRP) |
| (Lemma, PoS) | (Hamas_hamas, NP) | (HAMAS_hamas, NP) |
| (Synonym, PoS) | (view, NNS) | (opinion, NNS) |
| (Hypernym/Hyponym, PoS) | (perpetrator, NNS) | (offender, NNS) |

**Table 16** Pair of matches used in the Morphological Module

To the obvious use of the word-form reported in Table 16, the use of synonyms is also advisable, because they allow matches such as the pairs *believe-think* and *serious-dangerous* as illustrated in Example 72.

Example 72

HYP: *He said "I **believe** that the situation more difficult and complicated and **serious** which had been the several decades".*

REF: *He added "I **think** the situation is more difficult, complicated and **dangerous** than it has been for a number of decades."*

Likewise, hypernymy and hyponymy have also been considered as possible matches, since results obtained in the Lexical Module showed that they might be useful in a single-reference scenario.

On the other hand, there is some linguistic information used in the Lexical Module that has been avoided in the Morphological one: the partial lemma. This linguistic feature plays a role in the Lexical Module because it broadens the coverage of matches between lexical items as regards semantics. However, its use in combination with PoS tags does not seem relevant, since the Morphological Module tries to narrow the scope of the metric by dealing with information regarding inflectional morphology.

So as to see the relevance of each type of match, experiments were carried out both in a 4-reference and a single-reference scenario (see Table 17 and Table 18, respectively).

| Linguistic Features | Pearson Correlation |
|---|---|
| Word-form, PoS | 0.5902 |
| Word-form + lemma, PoS | 0.6011 |
| Word-form + lemma + synonymy, PoS | **0.6519** |
| Word-form + synonymy, hypern/hypo, PoS | 0.6503 |

**Table 17** Influence of each type of match in a 4-reference-scenario

| Linguistic Features | Pearson Correlation | | | |
|---|---|---|---|---|
| | **Ref. 1** | **Ref. 2** | **Ref. 3** | **Ref. 4** |
| Word-form, PoS | 0.5293 | 0.5191 | 0.5494 | 0.5316 |
| Word-form + lemma, PoS | 0.5533 | 0.5253 | 0.5478 | 0.5270 |
| Word-form + lemma + synonymy, PoS | 0.5908 | **0.5715** | 0.5931 | 0.5479 |
| Word-form + synonymy, hypern/hypo, PoS | **0.5911** | 0.5672 | **0.5959** | **0.5479** |

**Table 18** Influence of each type of match in a single-reference scenario

In line with those results obtained in the Lexical Module, the use of synonyms widens the coverage of the lexical items both in a 4 and a single-reference scenarios. Regarding the use of hypernyms and hyponyms in combination with PoS, similar results are obtained in a 4 reference scenario. Likewise, and in line with results reported in the Lexical Module section, their use is advisable when only one reference is available.

As regards PoS tags matches in isolation, correlation with human judgements on adequacy are low (0.4948) since these features are more related to the grammaticality of a segment.

### 6.3.2 Morphological Module vs. Lexical Module

Now that the morphology matches have been analysed and weighed, a comparison between the Lexical Module and the Morphological Module should be made. Our hypothesis that the Morphological Module was not as useful as the Lexical Module to assess adequacy has been confirmed by correlating the scores obtained by each module separately with the human judgements. As shown in Table 19, the Lexical Module with all linguistic features available shows a better correlation than the Morphological Module (0.7438 and 0.6519, respectively).

| Lexical Module | | Morphological Module | |
|---|---|---|---|
| **Linguistic Features** | **Pearson Cor.** | **Ling. Features** | **Pearson C.** |
| Word-form | 0.5766 | Word-form, PoS | 0,5902 |
| Word-form + Lemma | 0.7212 | Word-form, Lemma, PoS | 0.6011 |
| Word-form + Lemma + Synonymy | 0.7399 | Word-form + Lemma + Syn, PoS | **0,6519** |
| Word-form + Lemma + Synonymy + Partial lemma | **0.7438** | | |

**Table 19** Lexical Module vs. Morphological Module

It must be highlighted that there is a slight difference in favour of the Morphological Module when the only lexical relation used is the word-form. In the Morphological Module, this feature combined with the PoS gets 0,5902, whereas the same feature used by the Lexical Module, and therefore not taking into account PoS gets 0,5766. Data has revealed that this difference is a consequence of a different PoS-tagging in the hypothesis and reference segments, as shown in Example 73.

Example 73

HYP: *The statement said that "talks were held in a climate of confidence and friendly relations between the two countries strength reflects a* ***general****-JJ and interior ministries in particular between."*

REF: *The statement noted that "The discussions took place in an atmosphere of mutual trust and friendship that reflects the solidarity of the existing relations between the two countries in* ***general****-NNS, and between the two Interior Ministries in particular."*

In the hypothesis segment, and due to the ill-formation of the last part of the segment, the position of the word *general* is being altered; thus occupying the position of a coordinated adjective premodifying the noun *ministries*. As a result, the automatic tagger identifies it as an adjective, whereas in the reference segment the same word is tagged as a noun.

It is obvious that the ill-formation of this last part of the hypothesis segment affects its meaning; thus, the adequacy judgement obtained. It seems to indicate therefore that when only the word-form is used, the use of PoS information can also be useful when assessing adequacy because it can give a clue to identify issues in morphosyntax and in the grammaticality of the sentence that negatively affect its meaning and prevents the segment from being understood.

Likewise, PoS information can also help adequacy as regards verb tense, as exemplified by the words *are* and *were* below.

Example 74

HYP: *"...although some **are**-VBP born in Netherlands."*

REF: *"...some of them **were**-VBD born in Holland."*

*Are* and *were* share the same lemma, thus they would make a positive match in the Lexical Module, although the former is a present tense and the latter a past tense. However, when using the Morphological Module these two words will not match due to the difference in the PoS. Although the influence in meaning is not crucial, a different verb tense may also be taken into account when assessing adequacy.

### 6.3.3 Summing Up

This section has dealt with the use of PoS information in combination with lexical matches to assess adequacy. The same type of matches as in the Lexical Module have been used, with the exception of the partial lemma. Moreover, the use of hypernyms and hyponyms has also been tackled, obtaining similar results to those obtained in the Lexical Module.

As regards the use of weights, the same weights assigned in the Lexical Module have been used in the Morphological Module, although as stated in section 6.2.3 more data is required to reach final weights.

Finally, as expected, the correlation with human adequacy judgements is lower than that obtained by the Lexical Module. This indicates that such a module might not be useful to assess the adequacy of a segment.

Next, we continue exploring the use of linguistic information by testing the Dependency Module.

### 6.4 Dependency Module

The Dependency Module relies on dependency-relation matches in order to determine how similar two segments are. As explained in Chapters 3 and 5, dependency relations are established by the Standford parser (see sections 3.2.2.1 and 5.2.3 for further details). What is interesting about the use of dependency matches is that they can account for word order changes inside the clause (see Figures 12 and 13), as exemplified below:

Example 75

HYP: *After a meeting Monday night with the head of the Egyptian intelligence chief Omar Suleiman [Adj-Time] Haniya [Subj] said [Verb]...*

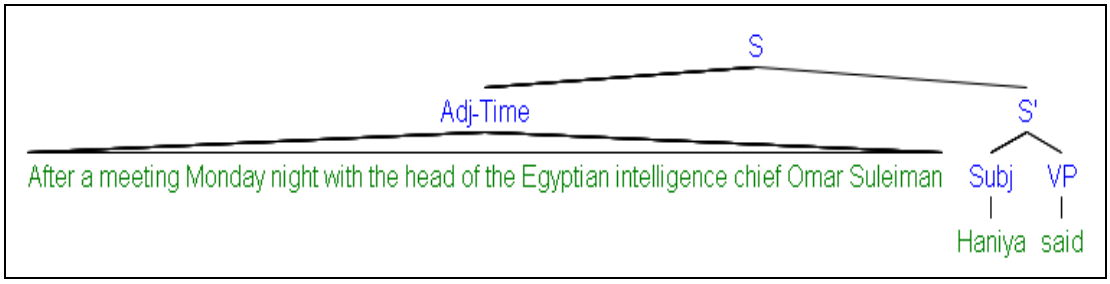



**Figure 12** Syntactic Tree corresponding to the hypothesis segment in Example 75

REF: *Haniya [Subj] said [Verb], after a meeting on Monday evening with the head of Egyptian Intelligence Omar Suleiman [Adj-Time]...*

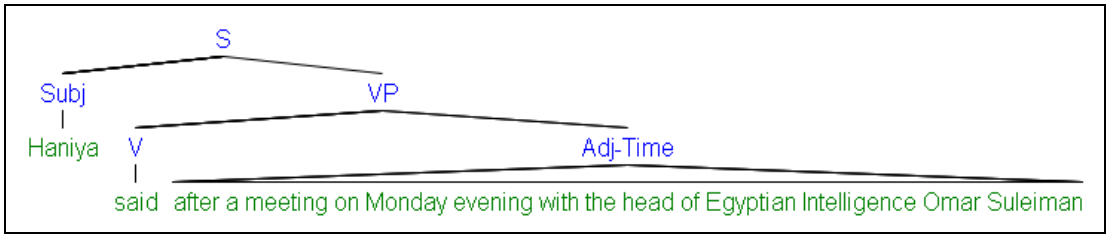



**Figure 13** Syntactic Tree corresponding to the reference segment in Example 75

The adjunct of time occupies an initial position in the hypothesis segment, followed by the subject and the verb (see Figure 12). On the other hand, the reference segment follows a more canonical word order: Subject-Verb-Adjunct (see Figure 13). As

reported in Table 20, despite this different word order, the meaning of the hypothesis segment is not affected and is absolutely comprehensible and comparable to the reference segment.

| Hypothesis | Reference | Match |
|---|---|---|
| prep_after(said,meeting) | prep_after(said,meeting) | Exact match |
| nsubj(said,Haniya) | nsubj(said,Haniya) | Exact match |

**Table 20** Matching of dependency triples exemplifying the different word order of the Adjunct of time in Example 75

The current section is organised as follows: in 6.4.1 the different types of matches used are described and linguistic data supporting their use is provided; in 6.4.2 the weights assigned to each type of match are stated; in 6.4.3 the dependency labels used and their corresponding weights are presented; in 6.4.4 the language-dependent rules which have been implemented to assess Arabic-English are covered; finally, in 6.4.5 a summary of the use of this module is provided.

### 6.4.1 Types of Matches

As described in section 5.2.3, the dependency matches established are the **Exact match**, where there is no difference between the hypothesis triple and the reference triple; **No_label match**, where the head and the modifier of both hypothesis and reference triple coincide but the dependency label is different; **No_mod match**, where the modifier is different but the label and head coincide; and **No_head match**, where both label and modifier coincide but the head is different.

First, experiments were carried out so as to check which types of matches were the most appropriate. In order to check their appropriateness, all dependency tags as well as each type of match were given the maxium weight and the scores provided by the metric were correlated with human judgements. Results obtained are shown in Table 21.

| Type of Match | Pearson Correlation |
|---|---|
| Exact match | 0.6505 |
| Exact + No_label | 0.6732 |
| Exact + No_label + No_mod | 0.6697 |
| Exact + No_label + No_mod + No_head | **0.7274** |

**Table 21** Influence of each type of match in the Dependency Module in a 4-reference scenario

Results reported in Table 21 show that the best correlation is obtained when all types of matches are applied (0.7274). However, apart from the impact of the Exact match, the highest impact is achieved when adding the No_head match, whereas the No_mod match slightly worsens the correlation with human judgements. In order to confirm those results, similar experiments were performed using a single reference (see Table 22).

| Type of Match | Ref. 1 | Ref. 2 | Ref. 3 | Ref. 4 |
|---|---|---|---|---|
| Exact | 0.5956 | 0.5883 | 0.5594 | 0.4852 |
| Exact + No_label | 0.6208 | 0.6007 | 0.5744 | 0.4889 |
| Exact + No_label + No_mod | 0.6069 | 0.5991 | 0.5758 | 0.4746 |
| Exact + No_label + No_mod + No_head | **0.6686** | **0.6620** | **0.6436** | **0.5583** |

**Table 22** Influence of each type of match in the Dependency Module in a single-reference scenario

As shown in Table 22, the performance of the metric is similar as regards the impact of the different types of match. The type of match that has the strongest influence is that obtained adding the No_head match, whereas the match with the weakest influence is the No_mod match. Actually, the No_mod match causes a decrease in the correlation with human judgements when both 1-reference and 4-reference scenarios are considered.

Next, each type of match is explained in detail with exception of the Exact match, due to its obvious use.

#### 6.4.1.1 No_label Match

The addition of the No_label match helps to identify different syntactic structures that can be used to convey the same meaning, as illustrated in Examples 76, 77 and 78.

Example 76

HYP: ...[*Russian President Putin (NP-Subj) invited (VP-Verb) HAMAS leader (NP-Od) to visit....* ***(X-Compl)***] ***Clause***

REF: ...[*Russian* ***President*** *Vladimir* ***Putin****'s (NP-Gen) invitation (N-Head) to the Hamas leadership (PP-Post-mod) to visit... (No-fin Cl-Post-mod)*] ***NP***

In Example 76, the syntactic structure of the hypothesis segment is a clause showing the clause pattern SVOXCompl, whereas the reference segment is realised by a NP, whose head is premodified by a NP-genitive and postmodified by a PP and a Non-finite clause. Although syntactically different, both structures express the same meaning and the Dependency Module is capable of identifying this similarity beyond the different syntactic structure thanks to the No_label match, as reported in Table 23.

| Hypothesis | Reference | Match |
|---|---|---|
| nn(Putin,President) | nn(Putin,President) | Exact match |
| **nsubj(invited,Putin)** | **poss(invitation,Putin)** | No_label match |
| nn(leader,HAMAS) | nn(leadership,Hamas) | Exact match |
| dobj(invited,leader) | **prep_to(invitation,leadership)** | No_label match |
| aux(visit,to) | aux(visit,to) | Exact match |

**Table 23** Example of Exact matches and No_label matches corresponding to Example 76

As for Example 77, the hypothesis segment is realised by a NP, whose head *delegation* is postmodified by a PP *of Moroccan police*. The reference segment is also realised by a NP whose head is also *delegation*, but premodified by a NP *Moroccan police*. The syntactic realisation of both segments is different, although their meaning is equivalent.

Example 77

HYP: ... *a delegation of Moroccan police...*

REF: ...*a Moroccan police delegation...*

Thanks to the use of No_label match (see Table 24), this similarity in meaning is captured.

| Hypothesis | Reference | Match |
|---|---|---|
| det(delegation,a) | det(delegation,a) | Exact match |
| **prep_of(delegation,police)** | **nn(delegation,police)** | No_label match |
| amod(police,Moroccan) | amod(police,Moroccan) | Exact match |

**Table 24** Example of Exact matches and No_label matches corresponding to Example 77

In Example 78, both segments start with an adjunct of time, however, the adjunct in the hypothesis segment is realised by a PP, whereas that in the reference segment is realised by a NP.

Example 78

HYP: *After four days [Adj-PP], four other churches burned down in three counties neighbouring Alabama West.*

REF: *Four days later [Adj-NP], four more churches in three other neighbouring counties to the west of Alabama were burned down.*

Despite showing a different surface realisation, both adjuncts share the same meaning, which is identified by adding the No_label match (see Table 25).

| Hypothesis | Reference | Match |
|---|---|---|
| num(days,four) | num(days,four) | Exact match |
| **prep_after(burned,days)** | **tmod(burned,days)** | No_label match |
| NO MATCH | advmod(later,days) | No match |
| nsubj(burned,churches) | nsubjpass(burned,churches) | No_label match |
| admod(burned,down) | prt(burned,down) | No_label match |
| NO MATCH | auxpass(burned,were) | No match |

**Table 25** Example of Exact matches and No_label matches corresponding to Example 78

Using this type of match favours the identification of matchings that are correct as regards adequacy but may pose a problem to the fluency of the sentence, as illustrated in Examples 78 and 79.

In Example 78, the reference segment contains a passive structure *were burned down*; however, in the hypothesis segment the *be* passive is missing, leading to a mismatch between hypothesis and segment (see Table 25). A close analysis of the hypothesis sentence, though, reveals that this mismatch does not affect the meaning of the segment because the semantic link between *churches* and *burned down* is not broken, actually the reader can perfectly understand that *4 churches were burned down*, even without the *be* passive. The use of the No_label match helps the metric to account for such a semantic relation.

Example 79 also exemplifies that the use of the No_label match may cause some problems as regards the grammaticality of a sentence, although this might not be an obstacle to understand its meaning.

Example 79

HYP: *This series of events in the Beba province [Subj] started [Verb] burning five churches [XCompl] in the 3rd February [Adj]*

REF: *The series of incidents [Subj] began [Verb] with the burning of five churches in Bibb County on February 3rd [Obl]*.

It is clear that, according to both hypothesis and reference segments, the burning of five churches was the beginning of a series of events. In the reference, the verb *began* is followed by an Oblique introduced by preposition *with*, which is followed by the nominalisation of a gerund *the burning of*. On the other hand, in the hypothesis segment, the verb *started* is followed by an X-Complement, headed by the gerund *burning*. Although such a syntactic structure may sound a bit odd in English, it does not prevent the reader from understanding this sentence. Therefore, the use of No_label match is also justified in this case, as shown in Table 26.

| Hypothesis | Reference | Match |
|---|---|---|
| nsubj(started,series) | nsubj(began,series) | Exact match |
| **xcomp(started,burning)** | **prep_with(began,burning)** | **No_label match** |
| **dobj(burning,churches)** | **prep_of(burning,churches)** | **No_label match** |
| num(churches,five) | num(churches,five) | Exact |

**Table 26** Example of Exact matches and No_label matches corresponding to Example 79

#### 6.4.1.2 No_mod Match

According to the correlation with human judgements (see Table 21 and Table 22), the addition of a match that only focuses on the similarity between label and head, disregarding the modifier, worsens the performance of this Module. Beyond correlation, if the data is analysed in detail, it seems that this match should not be disregarded in all cases, since it helps in cases like:

a) It helps to match two lexical items that do not share any of the linguistic features covered in the Lexical Module and that should me matched. That is the case of *illegal* and *clandestine* in Example 80.

Example 80

HYP: *...in the field of combating **illegal** immigration and drug smuggling.*

REF: *...in the areas of fighting **clandestine** immigration and drug smuggling.*

| Hypothesis | Reference | Match |
|---|---|---|
| det(field,the) | det(areas,the) | Exact match |
| prepc_of(field,combating) | prepc_of(areas,fighting) | Exact match |
| **amod(immigration,illegal)** | **amod(immigration,clandestine)** | **No_mod match** |
| dobj(combating,immigration) | dobj(fighting,immigration) | Exact match |
| nn(smuggling,drug) | nn(smuggling,drug) | Exact match |
| conj_and(immigration,smuggling) | conj_and(immigration,smuggling) | Exact match |

**Table 27** Dependency triples match corresponding to Example 80

The words *illegal* and *clandestine* although semantically related are not captured as equal items by the Lexical Module, mainly because they are not considered synonyms nor hypernyms in WordNet, neither share the same lemma or partial lemma. However, it is undeniable that they convey a similar meaning in this context and the No_mod match helps in dealing with it (see Table 27).

b) In addition, this match also helps in understanding the meaning of a sentence or part of a sentence, even if some parts are missing or wrong prepositions are being used, as reported in Example 81. Although the NP *a statement* is missing and the preposition *in* is incorrect in the hypothesis segment, the use of the No_mod match (see Table 28) prevents this omission and the incorrect preposition from being penalised, and helps to improve the correlation with the human judgement assigned to this segment (a score of 5) that increases from 4 to 5 with the addition of this match.

Example 81

HYP: *The minister said in an Israeli radio...*

REF: *The minister said in **a statement** on Israeli radio...*

| Hypothesis | Reference | Match |
|---|---|---|
| det(minister,the) | det(minister,the) | Exact match |
| nsubj(said,minister) | nsubj(said,minister) | Exact match |
| det(radio,an) | NO MATCH | No match |
| amod(radio,Israeli) | amod(radio,Israeli) | Exact match |
| **prep_in(said,radio)** | **prep_in(said.statement)** | **No_mod match** |

**Table 28** Dependency triples match corresponding to Example 81

On the other hand, such a match also has its drawbacks, the most important one being incorrect matches affecting key syntactic relations in the sentence, as illustrated below.

Example 82

HYP: *...saying that **it " imagine** themselves Hitler".*

REF: *...deeming that "**she thinks** she is Hitler."*

In this example, the translation engine fails in translating the subject of verb *imagine*, which is a serious mistake since it affects one the main syntactic functions in the sentence. By using the No_mod match, the metric matches *it* in the hypothesis segment with *she* in the reference segment as the subject of the verb (see Table 29), which is absolutely misleading. This is one of the reasons why the score obtained by this hypothesis segment moves from 3.6 up to 4.8, leading to a decrease in the correlation with the human judgement assigned to this sentence: 4.

| Hypothesis | Reference | Match |
|---|---|---|
| complm(imagine,that) | complm(thinks,that) | Exact match |
| nsubj(imagine,it) | nsubj(thinks,she) | No_mod match |
| ccomp(saying,imagine) | ccomp(deeming,thinks) | No_head match |

**Table 29** Dependency triples match corresponding to Example 82

The bad effect of this module can be balanced later by the language-dependent restrictive rules (see section 6.4.4) or by combining this module with the N-gram Module.

#### 6.4.1.3 No_head Match

The No_head match is the last type of match and probably the most controversial. Although some researchers such as Owczarzak et al. (2007a/b) used it in their experiments, most recent studies (He et al. 2010) have disregarded its use, apparently because it did not correlate well with human judgements. Our experiments seem to confirm the former's approach because it significantly improves the correlation with human judgements (see Tables 21 and 22). What is more, from a linguistic perspective the use of this match results in a more flexible Dependency Module and leads to a better performance of VERTa to evaluate adequacy. A detailed analysis of the data used has revealed the following:

a) The No_head match – similar to the No_mod match – helps to link lexical items that are semantically related but are not matched by the Lexical Module as shown in Example 83, where the words *said* and *added* are not considered as synonyms by the Lexical Module and would be disregarded as a possible match, but for the No_head match that takes them into account (see Table 30).

Example 83

HYP: *He **said** "I believe that the situation more difficult and complicated and serious which had been the several decades".*

REF: *He **added** "I think the situation is more difficult, complicated and dangerous than it has been for a number of decades."*

| Hypothesis | Reference | Match |
|---|---|---|
| **nsubj(said,He)** | **nsubj(added,He)** | **No_head match** |
| _( TOP,said) | _( TOP,added) | No_mod match |
| **punct(believe,")** | **punct(added,")** | **No_head match** |
| nsubj(believe,I) | nsubj(think,I) | Exact match |
| **ccomp(said,believe)** | **ccomp(added,think)** | **No_head match** |

**Table 30** Dependency triples match corresponding to Example 83

By means of the No_head match, the semantic relations between the verb and the subject, as well as the verb and the clause complement, are identified. This results in an

improvement of the score assigned by the metric to this segment, which moves from 2.7 up to 4.5, closer to 6, the human judgement assigned.

b) It helps to the general understanding of the sentence or part of a sentence. In Example 84, the use of the No_head match links the relative clause introduced by a relative pronoun *who* in the hypothesis segment and *which* in the reference, with their respective referents *HAMAS* and *movement*, stating that these words refer to the same entity (see Table 31). Such a relation would be lost if only Exact and No_label matches were used.

Example 84

HYP: *Cairo 7-2 (AFP) – announced parliamentary bloc, chairman of the Islamic Resistance Movement (HAMAS), Ismail Haniya,* ***that HAMAS who won the legislative elections in late January will offer*** *an official Fatah Movement led by Mahmoud Abbas to participate in the next government.*

REF: *Cairo 2-7 (AFP) – The leader of the parliamentary bloc of the Islamic Resistance Movement (Hamas) Ismail Haniya announced* ***that the movement, which won the legislative elections at the end of January, will formally invite*** *the Fatah movement led by Mahmoud Abbas to participate in this government.*

| Hypothesis | Reference | Match |
|---|---|---|
| **complm(offer,that)** | **complm(invite,that)** | **No_head match** |
| nsubj(offer,HAMAS) | NO MATCH | No match |
| **rcmod(HAMAS,won)** | **rcmod(movement,won)** | **No_head match** |
| amod(elections,legislative) | amod(elections,legislative) | Exact match |
| dobj(won,elections) | dobj(won,elections) | Exact match |
| **aux(offer,will)** | **aux(invite,will)** | **No_head match** |

**Table 31** Dependency triples match corresponding to Example 84

c) It helps to match different structures that convey the same meaning and that use different lexical items, as stated in Example 85. The No_head match links the conjunction *that* to the verb heading the that-clause, *conduct* and *visit* respectively (see Table 32). In addition, it also helps to link these verbs to the non-finite verb *to put*

heading the X-complements. Although different words are used, the No_head match is flexible enough to account for such a syntactic similarity.

Example 85

HYP: *...adding **that** a delegation of Moroccan police will **conduct** within the next month visit to France **to put** the finishing touches to the framework agreement on cooperation in this area"*

REF: *...said **that** a Moroccan police delegation would **visit** France next month **to put** the final touches to the cooperation agreement in this area.*

| Hypothesis | Reference | Match |
|---|---|---|
| **complm(conduct,that)** | **complm(visit,that)** | **No_head match** |
| det(delegation,a) | det(delegation,a) | Exact match |
| xsubj(put,delegation) | xsubj(put,delegation) | Exact match |
| amod(police,Moroccan) | NO MATCH | No match |
| prep_of(delegation,police) | nn(delegation,police) | No_label match |
| **xcomp(conduct,put)** | **xcomp(visit,put)** | **No_head match** |

**Table 32** Dependency triples match corresponding to Example 85

In conclusion, both from a linguistic point of view and correlation with human judgements, the No_head match has proved effective.

### 6.4.2 Match Weights

Once the 4 types of dependency matches were analysed, not only taking into account their correlation with human judgements but also the linguistic analysis carried out, our next step was to work on the importance that each match should have. The first studies on the use of dependencies to assess the quality of machine translation did not use different parameters depending on the type of match; in other words, all matches had the same importance. However, most recent studies (He et al. 2010) have introduced different parameters depending on the type of match. Inspired by this tendency, our research has also studied if assigning different weights to each type of match had any influence on the results obtained.

According to the results obtained in the previous section (6.4.1), it is obvious that full credit must be assigned to the Exact match, whereas the No_mod match should receive the lowest weight. This is not only due to the poor impact that this match has on the correlation with human judgements, but also to the questionable effect that has been reported in the linguistic analysis presented in this chapter. As regards the other two matches – No_label match and No_head match –, it seems that similar weight should be given to both of them. The No_label match shows that there is a relation between two triples that share the same head and modifier but differ in the type of dependency relation. It does not mean that the dependency relations identified are wrong, on the contrary, as shown in the linguistic analyses conducted, those triples can belong to two different valid syntactic structures that convey the same meaning. Thus, similar weights to that assigned to the Exact match should be used. As for the No_head match, it helps to match lexical items that are close in meaning but cannot be handled by the lexical relations established in the Lexical Module. Likewise, it also allows for linking expressions that, despite not being completely accurate, are still meaningful and can be understood. Therefore, although we consider that the weight provided cannot be the same as that assigned to the Exact or No_label match, it cannot be as low as the one assigned to the No_mod match.

The weights finally assigned depend on the previous remarks and also on how well they correlated with human judgements (see Table 33). As expected, the Exact match and No_label match were assigned the same weight, the No_head match was given a high weight but slightly lower than the first two, and finally, the No_mod match was given the lowest weight.

| Type of match | Weight | Correlation with Human Judgements |
|---|---|---|
| Exact | 1 | 0.7447 |
| No_label | 1 | |
| No_head | 0.9 | |
| No_mod | 0.7 | |

**Table 33** Weights assigned to each type of match and their resulting correlation with human judgements

All in all, as the amount of data used is rather small, these weights should only be considered a tendency, for larger data the weights should be readjusted.

### 6.4.3 Dependency Labels

Following He et al. (2010)'s approach, dependency relations are given different weights depending on their importance. Our first hypothesis was that more prominent relations, such as the dependency relation between the subject and the verb should receive a higher weight than that between a determiner and a noun, because from a linguistic point of view, the former is more relevant to assess the adequacy of the segment analysed than the latter. Therefore, dependency relations were initially organised into 3 levels and three different weights were assigned accordingly.

- TOP: dependency relations[37] affecting those constituents that depend on the verb, auxiliary verbs (both modal and non-modal), and copular verbs. [nsubj, dobj, aux, ccomp, csubjpass, rcmod, auxpass, nsubjpass, xsubj, cop, advcl, agent, appos, neg, parataxis, csubj, iobj, acomp, expl, attr, purpcl, xcomp, tmod, root]: **1**

- MID: dependency relations mostly covering modifiers inside the phrase [amod, nn, prep, prep_*[38], conj_*, conj, advmod, prt, mark, pobj, cc, infmod, rel, pcomp, prepc_*, abbrev, partmod, ref]: **0.7**

- LOW: dependency relations mostly related to punctuation marks, determiners and unlabeled constituents [dep, det, discourse, punct, complm, poss, num, number, predet, npadvmod, quantmod, possessive, measure, preconj, mwe, _]: **0.5**

Experiments were conducted in order to check whether those weights were appropriate, using all types of matches with their corresponding weights and correlating the score obtained with the human assessment. The result obtained was 0.7426, which seemed to correlate slightly worse than that obtained when all labels were given the same weight (0.7447). Several weight combinations were tried until the weights that correlated best with human judgements were found. To our surprise, the sets of tags had to be modified, since, according to our experiments, all dependency labels should receive the

[37] For a full list of dependencies refer to Appendix A.
[38] The Sandford parser provides collapsed dependencies, thus * stands for any prepositions, conjunctions or prepositional clausal modifiers (e.g. prep_of, prep_in).

same weight (1) except for *dep*, *det* and _[39] that were assigned 0.5. The correlation obtained after implementing those changes was 0.7474, which is not too far from the correlation obtained when all labels were given the same weight. This indicates that more data is needed in order to establish final parameters.

### 6.4.4 Rules

The Dependency Module also allows for implementing rules in order to capture similarity between different syntactic structures and constrain the grammatical context where some elements appear. These rules are language-dependent and may vary depending on the source and target language and the type of evaluation performed. Thus previous linguistic knowledge is required in order to implement and use them.

The data used to develop the MT evaluation metric has been translated from Arabic into English (see section 3.2.1.1 for further details). As regards the target language, some syntactically-different structures expressing the same meaning have been identified and several rules have been developed to cover them. These structures are applied at phrase and clause level, the former affecting modifiers inside the noun phrase and the latter valency alternations regarding the verb and its complements (please refer to section 5.2.3 for further details). Currently, these equivalent structures are already covered by the No_label match, which in our experiments has been assigned the maximum weight. However, as weights reported here are not final weights but just a tendency, these rules would be helpful if a different weight should be finally assigned to the No_label match because they would grant the maximum weight to this type of structures.

Finally, the Dependency Module also supports more restrictive rules. When performing the first linguistic analysis of the data used to develop the metric, we realised that the different word order of the subject and the object in Arabic and English was still an unsolved problem to some MT systems (see section 4.2.4). Some examples were found that show MT systems failed in translating the subject (see Examples 50 and 51 in section 4.2.5) and in reordering, as shown in Example 86.

[39] *det* stands for determiner; *num* stands for numeral and _ refers to those intermediate categories that help moving from standard dependencies to collapsed dependencies.

Example 86

HYP: *HAMAS leaders also discuss with Arab League General Secretary General* ***Amr Moussa*** *will* ***meet****.*

REF: *Hamas leaders will also* ***meet Amr Moussa****, the secretary-general of the Arab League, for discussions*

According to our set of matches, this pair of triples would be covered by the No_label match (see Table 34), resulting in a completely wrong match. In order to avoid this issue, a set of restrictive rules on the subject (see section 5.2.3 for further details) have been applied.

| Hypothesis | Reference | Match |
|---|---|---|
| nsubj(meet,Moussa) | dobj(meet,Moussa) | No_label match |

**Table 34** Dependency triples match corresponding to Example 86

The use of this set of rules has slightly increased the correlation with human judgements from 0.7474 up to 0.7523.

### 6.4.5. Summing Up

This module contains information based on the Lexical Module matches, in other words, it also uses word-form, lemma, synonymy and partial lemma information. This module relies on triples matches of four types: Exact match, No_label match, No_mod match and No_head match. These matches have been assigned different weights based on how well they correlate with human judgements. Final weights were: maximum weight (1) for the Exact and No_label match, 0.9 for the No_mod match and 0.7 for the No_head match. In addition, dependency categories have also been assigned different weights depending on how informative they are, thus most of the categories receive the maximum weight (1) except for det, num and _ that receive (0.5). Finally a set of language-dependent features have been added to restrict certain dependency relations. The Dependency Module has proved quite effective to evaluate the adequacy of a segment.

In the next section, a quite different type of module which involves less linguistic information is explored, the N-gram Module

### 6.5. N-gram Module

The N-gram Module is mainly in charge of controlling the word order of lexical items inside the sentence. Thus, its use seems particularly appropriate to assess the fluency of a segment, although it may also help to assess its adequacy, and we have studied this.

Section 6.5.1 explores different types of N-gram matches and the linguistic features that might be more appropriate to evaluate the adequacy of a segment; section 6.5.2 compares the performance of the N-gram Module to that of the Lexical Module; finally, 6.5.3 summarizes the use of the N-gram Module.

#### 6.5.1. N-gram Matches

The N-gram Module can work taking as a basis the lexical items, the combination of lexical items and PoS or just the PoS. As regards adequacy and after analysing the performance of the Lexical and Morphological Modules, our hypothesis was that calculating n-grams over lexical items was the best option since they can contain all the information provided by the Lexical Module (word-form, synonyms, lemma, partial lemma, hypernyms and hyponyms). In order to confirm our hypothesis, the following experiments were conducted:

a) N-grams over lexical items

b) N-grams over PoS

c) N-grams over combinations of lexical items and PoS

For each experiment, several n-gram lengths were considered in order to study which length correlated best with human judgements on adequacy:

a) From bigrams to sentence-length n-grams

b) Only bigrams

c) Bigrams and 3grams

d) Bigrams, 3grams and 4grams

Table 35 shows the correlation of human judgements with the results obtained by each experiment.

| N-grams Length | N-grams over Lexical Items | N-grams over PoS | N-grams over Lexical Items + PoS |
|---|---|---|---|
| **2grams to sentence-length grams** | 0.4109 | 0.4187 | 0.3921 |
| **2grams only** | **0.7017** | 0.5610 | 0.6477 |
| **2grams and 3grams** | 0.6792 | 0.5895 | 0.6293 |
| **2grams, 3grams and 4grams** | 0.6557 | 0.5959 | 0.6084 |

**Table 35** Correlation of the N-gram Module with human judgements on adequacy

As reported in Table 35, the results that correlate best with human judgements on adequacy are those obtained by computing N-grams over lexical items using a bigram-length (0.7017). This reinforces our hypothesis that lexical items are more appropriate than PoS to assess adequacy.

The results obtained also reveal that the shorter the length of the n-gram, the better the correlation with human judgements. In our experiments, this module correlated best with human judgements when only bigrams were used. On the other hand, the distance covering from bigrams to sentence-length grams is clearly more restrictive and therefore correlates worse with human judgements on adequacy. Thus, this might be taken into account when dealing with fluency.

### 6.5.2. N-gram Module vs. Lexical Module

When comparing results obtained by the N-gram Module with those obtained by the Lexical Module, it must be noticed that although there is a difference between them, it is not a remarkable difference (see Table 36).

| Module Used | Pearson Correlation |
|---|---|
| Lexical Module | **0.7469** |
| N-gram Module | 0.7019 |

**Table 36** Difference between Lexical Module and N-gram Module

The N-gram Module is based on the matches provided by the Lexical Module, which means that it does not only take into account word-forms but also other lexical relations linked to semantics. The use of rich semantic information is essential to obtain such good results. In addition, n-grams contribute to identify problems in word order that may result in the incomprehensibility of the hypothesis segment or part of the segment (see Example 87) or in a different meaning from that of the source sentence (see Example 88).

Example 87

HYP: ...***station*** *"TV"* ***television***...

REF: ..."NTV" **television station**...

Although there is a mistranslated word *NTV*, it is the incorrect word order of the head of the NP *station* preceding the pre-modifier *TV television* in the hypothesis segment that makes this part difficult to understand.

Example 88

HYP: ***[Rod Larsen] said*** *on* ***radio talking Norwegian*** *that*...

REF: ***[Rod Larsen]***, ***speaking*** *to the* ***Norwegian*** *NRK* ***radio*** *station* ***said*** *that*...

In the hypothesis segment, a reader may understand that Rod Larsen said something on the radio and he said it in Norwegian, whereas in the reference sentence it is clear that he addressed the Norwegian radio, although we do not know the language he used to do it. The Lexical Module matches the lexical items regardless of their order inside the sentence, so the score obtained by this part of the hypothesis segment is rather high (7 words out of 8 for Precision) as most of its words also appear in the reference segment. However, if we apply the N-gram Module we see that there is only one 2gram match (*Rod Larsen*), and therefore, the score obtained by this module drops dramatically.

On the other hand, the N-gram Module is sometimes an obstacle to get higher scores, because it is too restrictive. There are several cases where the use of this module is more an obstacle than help:

a) Equivalent expressions as illustrated in the example below.

Example 89

HYP: ***A delegation*** *of* ***Hamas****...*

REF: ***A Hamas delegation****...*

Semantically speaking, both expressions share the same meaning, but no n-gram match can be established.

b) Words with a heavy semantic weight are disregarded because they do not belong to any n-gram match, as illustrated in Example 90.

Example 90

HYP: ***Oslo*** *6-2* ***(AFP) – Terje*** *Rod Larsen* ***former*** *UN* ***envoy*** *to* ***Middle East believed*** *that* ***the situation in the area****...*

REF: ***Oslo*** *2-6* ***(AFP) – Terje*** *Roed Larsen, the* ***former*** *United Nations* ***Middle East envoy, considered the situation in the region****...*

According to the Lexical Module the matches between the hypothesis and reference segments are the following:

H: Oslo (AFP)–Terje former envoy Middle East believed the situation in the area

R: Oslo (AFP)–Terje former Middle East envoy considered the situation in the region

On the other hand, according to the N-gram Module, the matches between hypothesis and reference segments of Example 90 are:

HYP: [(AFP) – Terje] [Middle East] [the situation in the area]

REF: [(AFP) – Terje] [Middle East] [the situation in the region]

This example shows that the use of n-gram matches in isolation is too restrictive in terms of meaning, even if only bigrams are used, because it leaves out words which are semantically important such as *envoy* or *believe*. It is clear that, although the sentence formed using only those words matched by the Lexical Module is not fluent it is capable of keeping the meaning of the original segment.

c) The omission of function words such as determiners or prepositions leads to disregarding meaningful lexical items, as shown in Example 91.

Example 91

HYP: *After a meeting Monday night with the head of Egyptian intelligence chief Omar Suleiman Haniya said...*

REF: *Haniya said, after a meeting on Monday evening with the head of Egyptian Intelligence General Omar Suleiman...*

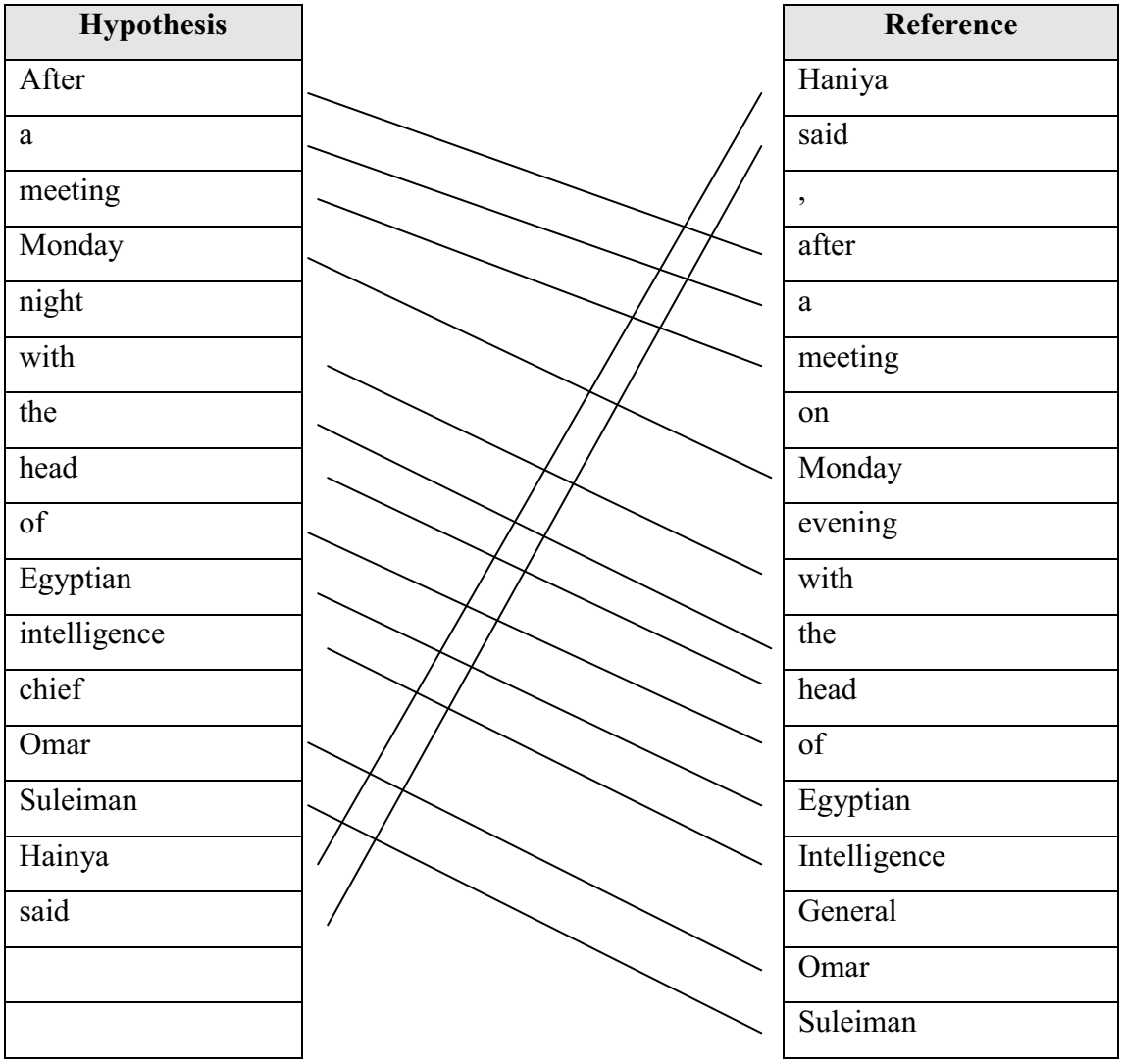
An important part of the meaning of the sentence is lost if the N-gram Module is applied in isolation: when the meeting took place. The word *Monday* that is captured by the lexical matches is disregarded by the n-gram matches because in the hypothesis segment the preposition *on* has been omitted. It is clear that the omission of such a preposition can affect the fluency of the segment but it does not prevent the understanding of its meaning.

Matches according to the N-gram Module:

HYP: [After a meeting] [with the head of Egyptian intelligence] [Omar Suleiman] [Haniya said]

REF: [Haniya said] [after a meeting] [with the head of Egyptian Intelligence] [Omar Suleiman]

Matches according to the Lexical Module:




### 6.5.3 Summing Up

Once the N-gram Module in isolation has been analysed, it can be concluded that in order to evaluate the adequacy of a segment, the N-gram Module has been more effective when based on lexical items, rather than PoS or the combination of lexical items and PoS. In addition, it has also been proved that the shorter the n-gram distance, the better the correlation with human judgements.

Next, the last type of module used to evaluate adequacy is analysed, the Semantic Module.

## 6.6 Semantic Module

As confirmed by the Lexical Module, semantics plays an important role in the evaluation of adequacy. This has also been confirmed by Lo and Wu (2010)'s work who report that their metric based on Semantic Roles outperforms other well-known metrics when adequacy is assessed:

> "Unlike the widely-used lexical and n-gram based or syntactic based MT evaluation metrics which are fluency-oriented, our results show that using Semantic Role labels to evaluate the utility of MT output achieve higher correlation with human judgments on adequacy". Lo and Wu (2010).

With this aim in mind other semantic features at sentence level have been explored and analysed in this section: NEs (section 6.6.1), Time expressions (6.6.2) and Sentiment analysis (section 6.6.3). Finally, section 6.6.4 deals with the combination of the components described above, and section 6.6.5 summarizes the use of the Semantic Module.

### 6.6.1 Named Entities (NEs)

As regards NEs, two components have been considered: NER (Named Entities Recognition) and NEL (Named Entities Linking). The NER component works similar to previous NE-based metrics (Reeder et al. 2011; Giménez 2008a) in the sense that it aims at capturing similarities between NEs in the hypothesis and reference segments. On the other hand the NEL component focuses only on those NEs that appear on Wikipedia, which allows for linking NEs regardless of their external form. Thus, *EU* and *European Union* will be captured as the same NE, as both of them are considered as the same organisation in Wikipedia.

#### 6.6.1.1 NER Component

As previously mentioned, the aim of this component is matching NEs in the hypothesis and reference segments. In order to identify NEs we use the Supersense Tagger (Ciaramita and Altun 2006). NER is extremely useful in order to check whether a NE has been translated correctly or it has not been translated at all. However, it must also be taken into account that one cannot expect this component to correlate well with human

jugdements on adequacy, as the latter are based on the adequacy of the whole segment and NER evaluates a partial aspect of the segment. Hence, correlation results at segment level are not particularly outstanding: **0.3387**.

In addition, a close analysis of the data reveals several issues that affect its performance:

- The tool used to recognize NEs considers all words that are written in capital letters as NEs, as shown in Example 92 (see Table 37).

  Example 92

  HYP: ***Rod Larsen****: **Middle East** "to prevent further **Baroud** barrels"*

  REF: ***Roed-Larsen****: **Middle East** a "**Powder Keg** with **Lit Fuse**"*

| NEs in Hypothesis | NEs in Reference |
|---|---|
| Rod Larsen [PER] | Roed-Larsen [PER] |
| Middle East [LOC] | Middle East [LOC] |
| Baroud [ORG] | Powder Keg [ORG] |
| | Lit Fuse [ORG] |

**Table 37** NEs recognized in the hypothesis and reference segments

  Neither *Baroud* nor *Powder Keg* nor *Lit Fuse* are real NEs, however the tool considers them as such because they are written in capital letters.

- NEs are not segmented properly, as illustrated in Example 93 (see Table 38).

  Example 93

  HYP: *Opening the leaders expressed their opposition to the government which is supposed to take part in that **Montenegro HAMAS** but that did not issue any official decision yet on the matter open.*

  REF: *A number of **Fatah** leaders expressed their opposition to participating in the government which is supposed to be formed by **Hamas**, but an official decision has not yet been issued by **Fatah** in this regard.*

| NEs in Hypothesis | NEs in Reference |
|---|---|
| Montenegro HAMAS [LOC] | Fatah [ORG] |
| | Hamas [ORG] |
| | Fatah [ORG] |

**Table 38** NEs segmented in the hypothesis and reference segments

In the hypothesis segment, the string *Montenegro HAMAS* should not be identified as a single NE, but as two different NEs *Montenegro*, a location, and *HAMAS* an organization. Unfortunately, the tool used fails in segmenting them.

- NEs that are not written in the same way do not match. That is the case of *Rod Larsen* and *Roed-Larsen* in Example 92, where these two NEs are identified as such but do not match because their external form is different. This is useful in order to detect bad translations, although on the other hand, this is a drawback when trying to match word-forms that refer to the same entity, such as *EU* and *European Union*. Such drawback, though, can be overcome by using the NEL component.

#### 6.6.1.2 NEL Component

The NEL component uses a graph-based NEL tool inspired by Hachey et al. (2011) which links NEs in a text with those in Wikipedia pages. Its aim is twofold: first, recognize bad translations of NEs, and second NE normalization, that is to say, identify different word-forms that refer to the same entity by means of checking if they are included in the same Wikipedia page. Such is the case of abbreviations (e.g. *USA* and *United States of America*), the use of someone's surname to refer to the whole name (e.g. *Merkel* and *Angela Merkel*) or the use of the position to refer to the person (e.g. *the German Chancellor* and *Angela Merkel*). It must be noticed though that as only NEs are being checked, the correlation with human judgements on adequacy will not be high, as only a partial aspect of the segment is being used. As regards our corpus, the correlation obtained is rather low: **0.1670**. This correlation is even lower than that obtained by the NER component due to the fact that those entities not contained in the wikipedia are disregarded, thus restricting the number of NEs matched. As shown in Example 94, not all NEs that are matched by the NER component (in bold), do necessarily match when the NEL component is applied.

Example 94

HYP: ***Rod Larsen****: **Middle East** "to prevent further Baroud barrels"*

REF: ***Rod Larsen****: The **Middle East** is "a barrel of gunpowder with a flaming wick."*

| NEs in Hypothesis | NEs Normalization | NEs in Reference |
|---|---|---|
| Middle East | Middle_East | Middle East |

**Table 39** NEs matched by the NEL component

Although the NEs identified in both segments are *Rod Larsen* and *Middle East*, the NEL component only shows one positive match (see Table 39), *Middle East*, since the Wikipedia webpage corresponding to *Rod Larsen* is *Rød Larsen*.

As previously mentioned, the NEL component helps in matching NEs regardless of their external form (if those NEs and their different physical representations are included in the Wikipedia), as illustrated in Example 95.

Example 95

HYP: ***Oslo** 6-2 (**AFP**) – **Terje Rod Larsen** former **UN** envoy to **Middle East** believed...*

REF: ***Oslo** 2-6 (**AFP**) – **Terje Rod Larsen**, the former **United Nations** **Middle East** envoy, considered...*

| NEs in Hypothesis | NEs Normalization | NEs in Reference |
|---|---|---|
| Oslo | Oslo | Oslo |
| Middle East | Middle_East | Middle East |
| AFP | Agence_France_Press | AFP |
| UN | United_Nations | United Nations |

**Table 40** NEs matched by the NEL component

In this example, the abbreviation of United Nations, *UN*, is used in the hypothesis segment whereas the full form is preferred in the reference. The NEL component links both of them to the corresponding Wikipedia webpage, thus allowing for a positive match (see Table 40).

In addition, the NEL component is also of great help when part of a NE has not been translated but the meaning is not affected, as shown in Example 96.

Example 96

HYP: ... *to **President George Bush**"*

REF: ... *to **American President George Bush**"*

| NEs in Hypothesis | NEs Normalization | NEs in Reference |
|---|---|---|
| President George Bush | George_W_Bush | American President George Bush |

**Table 41** NEs match by the NEL component

Although in the Hypothesis segment there is no reference to *American*, the meaning is not affected and our knowledge of the world allows us to infere that both NEs refer to George W. Bush, former president of the United States. By means of the NEL component those two chunks become a positive match (see Table 41), since *President George Bush* and *American President George Bush* refer to the same Wikipedia page.

Although the use of NEL has its weaknessess since it depends on the NEs gathered in the Wikipedia and it does not correlate well with human judgements on adequacy, it is helpful in order to link NEs with a different external form, which otherwise would be difficult to match.

### 6.6.2 Time Expressions (TIMEX) Component

This component matches temporal expressions in the hypothesis and reference segments regardless of their form. The tool used is the Stanford Temporal Tagger (Chang and Manning 2012) which recognizes and normalizes not only points in time but also duration.

Similar to the previous components this one does not show a high correlation with human judgements on adequacy since it only checks a partial aspect of the segment – those expressions referring to time. Thus, the correlation obtained is **0.2041**. However, as mentioned before, it is highly useful in order to match temporal expressions that could not be recognized as equivalents without this component (see section 4.2.1).

Next, some examples illustrating the usefulness of this component are presented:

- Equivalent Time Expressions of duration realised differently. In Example 97, a couple of equivalent Time Expressions conveying the meaning of duration are matched, regardless of the preposition introducing them.

  Example 97

  HYP: *The burning of churches 10* ***within 10 days***

  REF: *Ten Churches Burned Down* ***in 10 days***

| Time Expr. in Hyp. | TIMEX Normalization | Time Expr. in Ref. |
|---|---|---|
| Within 10 days | P10D: DURATION | In 10 days |

**Table 42** Time Expressions in the hypothesis and reference segments

  The expressions *within 10 days* and *in 10 days* are equivalent in meaning, but the PPs realising them are introduced by different prepositions. By means of the TIMEX component this similarity is captured (see Table 42).

- Equivalent Time Expressions of date realised differently. Example 98 illustrates different realisations used to express a date in English. The TIMEX component helps in considering them identical (see Table 43).

  Example 98

  HYP: *And this series of events started with burning five churches in Bib province* ***on third of February****.*

  REF1: *The series of incidents began with the burning of five churches in Bibb county* ***on February 3rd****.*

  REF2: *This series of incidents began with the burning of five churches in Bibb County* ***on the third of February****.*

| | Time Expressions | TIMEX Normalization |
|---|---|---|
| **Hypothesis** | on third of February | 2013-02-03:DATE |
| **Reference 1** | on February 3rd | |
| **Reference 2** | on the third of February | |

**Table 43** Time Expressions in the hypothesis and reference segments

The above expressions are different ways to refer to the same point in time. Thanks to TIMEX, equivalent moments in time are being identified and matched regardless of their different formats.

In addition, not only exact Time Expressions, such as those reported above, can be traced, but also rather vague Time Expressions realised by different syntactic structures, as shown in Example 99.

Example 99

HYP: ...*HAMAS who won the legislative elections* ***in late January****....*

REF: ...*the movement, which won the legislative elections* ***at the end of January****...*

| Time Expr in Hyp. | TIMEX Normalization | Time Expr. in Ref. |
|---|---|---|
| in late January | 2013-01: END | at the end of January |

**Table 44** Time Expressions in the hypothesis and reference segments

In this example an inexact date is expressed by two different syntactic structures. With the help of the TIMEX component both equivalent expressions are identified as a positive match regardless of their different syntactic realisation (see Table 44).

Unfortunately, this component does not always succeed in identifying similar expresions. Other formats are not considered such as those illustrated in Example 100, where the TIMEX tool fails in identifying the chunks in bold as Time Expressions, except for that in reference 2, and as a consequence, they are not linked as a positive match.

Example 100

HYP: *Oslo **6-2** (AFP) – ....*

REF1: *Oslo **2-6** (AFP)*

REF2: *Oslo, **February 6** (A.F.P.) - ...*

REF3: *Oslo **2/06** (A.F.P.) - ...*

REF4: *Oslo **6-2** (FB) - ...*

This is an example of the negative impact that the performance of the NLP tool used has on the performance of this component.

### 6.6.3 Sentiment Analysis Component

Part of the meaning of a sentence is directly linked to its positive or negative connotation. By means of Sentiment analysis we identify the contextual polarity of a text, in short, if a sentence is positive, negative or neutral. This feature is quite interesting since it goes beyond lexical semantics and the principle of compositionality, and contributes to identify one more characteristic in the whole meaning of a sentence. Actually, it transcends semantics itself since it tries to infer the real meaning behind a sentence.

In order to calculate the contextual polarity of a sentence, the dictionary strategy described in Atserias et al. (2012) has been used. The segment score is computed by:

- Adding 0.5 for each weak positive word.
- Adding 1.0 for each strong positive word.
- Substracting 0.5 for each weak negative word.
- Substracting 1.0 for each strong negative word.

For each sentence, a word score is aggregated to provide a score at sentence level, In order to obtain a final score between 0 and 1 the aggregation of the scores is divided by the number of words in the sentence. For example, segment (i) would obtain a score of

0.0041 as regards sentiment analysis because words like *danger*, *prevent*, *situation* and *mark* are considered as negative words; whereas words like *believed*, *Rod*, *envoy*, *Middle*, *day* are regarded as positive words. Thus the polarity of the sentence is closer to 0 (neutral).

(i) *Oslo 6-2 ( AFP ) - Terje Rod Larsen former UN envoy to Middle East believed that the situation in the area had not yet who is on its danger mark day today , the region , " to prevent further per Baroud " .*

Unfortunately, in the same line as the above mentioned components, since Sentiment analysis is just a partial aspect of sentence semantics, the correlation with human judgements of adequacy obtained is rather low: **0.1325**. In addition, the qualitative analysis of the data has shown that results obtained by the dictionary strategy described do not help much since most of the scores obtained tend to 0, in other words, to a neutral contextual polarity. This might be due to the domain of the data used, and indicates that this type of knowledge is more adequate for other domains such as product reviews or film reviews where a more subjective tone is used.

### 6.6.4 Combination of Components

Previous sections have shown that most of the information used in the Semantic Module separately does not correlate well with human judgements, although it has proved useful from a linguistic approach, except for the sentiment analysis. With the aim of exploring the overall influence of the Semantic Module, the above detailed components have been combined. Those components that were expected to show a stronger influence were the NER and TIMEX components, however final correlations indicate that neither NEL nor sentiment analysis should be entirely disregarded. Interestingly enough, the correlation obtained when all components are used is **0.3908**, better than that obtained by the NER component (**0.3387**), which had shown the best correlation when used individually.

### 6.6.5 Summing Up

In this section the Semantic Module has been introduced. This module uses information regarding NEs recognition, NEs linking, Time Expressions and Sentiment analysis. According to the experiments performed, this module shows a low correlation with

human judgements on adequacy, since only partial aspects of translation are considered, whereas human judgements assess the adequacy of the entire hypothesis segment. From the information contained in this module, the one that correlates best individually is that referring to NEs recognition, however, when all components are combined the correlation of the whole module improves.

It must also be highlighted that the performance of the Semantic Module will vary depending on the source and target language as well as the type of corpus evaluated, i.e. open or close-domain corpus. For example, if the source and target languages use different alphabets, the translation of proper nouns might pose a problem since they may have different spellings when transliterated into English (e.g. Arabic proper nouns transliterated into English). In addition if the domain is a closed-domain, proper nouns will be more controlled, restricted and standardized (e.g. car brands, car companies).

Until now each module in VERTa and its corresponding linguistic information have been explored individually. The following section analyses the combination of all modules.

### 6.7 Modules Combination

Once each module has been analysed separately, the most natural step seems to be the combination of those modules that proved to be the most effective to assess adequacy.

In terms of correlation with human judgements, each separate module obtained the correlations shown in Table 45.

| Module | Pearson Correlation |
|---|---|
| Lexical Module | **0.7438** |
| Morphological Module | 0.6519 |
| Dependency Module | **0.7523** |
| N-gram Module | 0.7017 |
| Semantic Module | 0.3908 |

**Table 45** Correlations with human judgements per module

Thus, according to correlations, those that showed a better performance to assess adequacy were the Lexical and Dependency Module. These results can also be justified

from a linguistic point of view. The Lexical Module uses semantic relations to match hypothesis and reference lexical items, thus lexical semantics is being considered. In addition, the Dependency Module helps in matching different syntactic structures that convey the same meaning (see section 6.4) and accounts for sentence and phrase compositionality.

In order to confirm the ideal combination of modules, a couple of experiments were conducted (see Table 46):

1. Experiment 1: all modules were combined and all of them were assigned the maximum weight.

2. Experiment 2: all modules were used and weights were automatically tuned.

| Modules Combination | Pearson Correlation |
|---|---|
| All modules - maximum weight | 0.6176 |
| Lexical, Morphological, Dependency, N-gram and Semantic Modules (weights being 0.47, 0, 0.43, 0.05 and 0.05 respectively) | **0.7816** |

**Table 46** Modules combination

Automatic tuning confirms that the combination of the Lexical and Dependency Modules proves to be the most effective to assess adequacy, although neither the N-gram Module nor the Semantic Module should be omitted. The Lexical Module has the strongest influence (0.47), followed by the Dependency Module (0.43). This is not surprising since the Lexical Module covers lexical semantic relations, as mentioned above. In addition, corroborating one of our hypothesis, the Dependency Module also proves effective since it accounts for dependency relations, which are somewhere between syntax and semantics. Although n-gram information is usually connected to the grammaticality of a sentence, it cannot be completely disregarded, mainly in a language such as English, which shows a strict word order. As explained in section 6.5, wrong word order may lead to the incomprehensibility of the hypothesis segment or part of it. Nonetheless, it must be highlighted that a stronger use of the N-gram Module may also result in a too restrictive performance of the metric leading to a rather fluency-oriented evaluation; thus, as confirmed by the automatic tuning, the weight assigned to this

module should be rather low (0.05). Likewise, the influence of the Semantic Module, containing information regarding NEs, Time expressions and Sentiment analysis, is also rather small because it accounts for very partial aspects of the translated sentence. From a linguistic perspective, the use of the Morphological Module was not appropriate either, due to the language analysed (which is also confirmed by the automatic weight assigned to this module (0)). English does not show a rich inflectional morphology, thus this should not imply a major problem to MT systems. Actually, the instances of morphology issues (e.g. verb tense or number in nouns) in the corpus analysed were rather low. In addition, morphology seems to be more relevant when evaluating the fluency of a segment (see Chapter 7) than its adequacy or meaning.

A closer analysis of these results and of those examples that most benefit from this combination shows the following:

- Firstly, the Dependency Module infers relations that might be disregarded if only the Lexical Module was taken into account, as illustrated in Example 101[40].

Example 101

HYP: ***He*** *said* ***"that all these positions unfair to*** *the right* ***people,*** *US, and* ***we*** *now* ***possess*** *an* ***Islamic*** *or the* ***Palestinians and Arabs options".***

REF: ***He*** *added,* ***"****We emphasized* ***that all these positions*** *are* ***unfair to*** *our* ***people*** *but that* ***we have alternative Palestinian, Arab****,* ***and Islamic*** *resources****."***

In the hypothesis segment the copula verb *are* is missing, however, the meaning is not affected, and a potential reader could still infer that *all these positions are unfair*. If only the Lexical Module was taken into account, this no-match would penalise the hypothesis segment; however, if the Dependency Module is used, this relation between the subject *positions* and the Subject Complement (Cs) *unfair* is still preserved as shown below (see Table 47), where the analysis of the hypothesis segment accounts for a dependency relation between *position* and *unfair* although the type of relation cannot be established due to the missing copula verb. In addition, by means of the Dependency Module the clauses introduced by *and* and *but* in the hypothesis and reference segments,

[40] Lexical module matches in bold and N-gram module matches underlined.

respectively, can also be connected to the previous clause. Although the meaning of both connectors is clearly different, it does not seem to affect the meaning of the whole sentence. Finally, the Dependency Module also accounts for the last part of the sentence which shows a different word order as well as a clearly disfluent clause *we now possess an Islamic or the Palestinians and Arabs option*s, but whose meaning can still be understood.

| Hypothesis | Reference | Match |
|---|---|---|
| dep(unfair, positions) | nsubj(unfair, positions) | No_label match |
| conj_and(unfair, possess) | conj_but(unfair, have) | No_label match |
| dobj(possess, Islamic) | dobj(have, Palestinian) | No_head match |
| conj_or(Islamic,Palestinians) | NO MATCH | No match |
| conj_and(Palestinians,Arabs) | conj_and(Palestinian,Arab) | Exact match |
| dep(Palestinians,options) | amod(Palestinian,alternative) | No_label match |

**Table 47** Dependency matches corresponding to Example 101

- Secondly, the Dependency Module accounts for matches between different syntactic structures that express the same meaning (see Examples 75-79 in section 6.4).

- Thirdly, the N-gram Module helps to identify word order issues that may affect the understanding of the hypothesis segment (see Example 87 and Example 88 in section 6.5).

- Finally, the Semantic Module covers linguistic features regarding NEs, equivalent Time Expressions realised differently and Sentiment analysis (see section 6.6).

### 6.8 Findings on Adequacy

In this chapter, experiments for each module in VERTa have been conducted with the aim of checking the appropriateness of the linguistic features contained in the modules, as well as the modules themselves, to evaluate the adequacy of a segment. Experiments on a per-module basis have yielded the following findings:

**Lexical Module**. This module relies on traditional matches, namely word-form, synonymy, lemma and partial lemma. In addition, hypernyms and hyponyms were also used in order to prove our hypothesis that other semantic relations could be appropriate

to evaluate adequacy. Although correlations with human judgements on a 4-reference scenario decrease when these features are used, they tend to show a more positive influence when only one reference is available, as they seem to widen the coverage at lexical level. Likewise, a linguistic analysis of the results obtained has also corroborated this tendency. Therefore, although final conclusions cannot be drawn, hypernymy/hyponymy relations is a new feature that should not be totally disregarded in MT evaluation, especially when only one reference is available. Besides, following METEOR, the linguistic features used have been assigned different weights according to their importance in terms of meaning. Results confirmed that the use of such parameters leads to a slight increase in the correlation with human judgements. However, the data used in these experiments only show a tendency, a larger and different collection of data would be needed in order to obtain final weights.

**Morphological Module**. The Morphological Module, using PoS information, does not help in the evaluation of adequacy in English since it is too restrictive and thus more appropriate to evaluate the fluency of a segment, as it will be shown in Chapter 7. This corroborates our hypothesis that, in general, PoS information is not crucial to evaluate adequacy, at least in English. In addition, hypernyms and hyponyms have obtained similar results to those in the Lexical Module, which seems to confirm, once more, that this lexical semantic information might be helpful when only one reference is available.

**Dependency Module**. The Dependency Module has proved effective to establish similarities between different syntactic structures conveying the same meaning. Four different types of matches with different weights assigned were implemented. From those different matches, the Exact match, the No_label match and the No_head match improve the correlation with human judgements. In addition, although the No_mod match does not show strong correlation with human judgements, linguistic analysis has shown that it should not be completely disregarded because it allows for a broader coverage of syntactic relations.

The rather flexible matches in this module have made it particularly appropriate to evaluate adequacy. This corroborates our initial hypothesis that depending on how syntactic information is used, it can also be useful to evaluate de adequacy of a segment. We have proved that dependency relations can also account for the adequacy of a

segment, since they reflect the compositionality of the sentence and are somewhere between syntax and semantics (see section 1.1). In addition, the importance of the dependency labels has been explored and finally all of them were assigned the same weight – thus, being considered equally important – except for 3: det, num and _. This might indicate that most of the types of dependency relations are relevant when evaluating adequacy. However, more data would be needed to corroborate this tendency.

**N-gram Module**. This module has also proved effective to evaluate adequacy, although less than the Dependency and Lexical Modules. It must be highlighted that best results are obtained when n-grams are calculated over lexical matches, therefore using information regarding word-form, synonyms, lemma and partial lemma. In addition, the most appropriate n-gram length seems to be bigrams, since longer n-grams are too restrictive.

On the other hand, the use of the N-gram Module in isolation is too restrictive and may penalise sentences that are good as regards adequacy, however, it is also useful to control wrong word order that may lead to the incomprehensibility of the sentence or to a different meaning from that in the reference sentence. Thus, it is helpful if combined with another module which accounts for those lexical items disregarded by the N-gram Module, i.e. the Lexical Module.

**Semantic Module**. The Semantic Module uses information on NEs recognition, NEs linking, Time Expressions and Sentiment analysis. In general this module in isolation does not correlate well with human judgements on adequacy since the above mentioned features only assess very partial aspects of the segment. However, regardless of how well/badly the module correlates with human judgements, checking partial aspects of the segments translated, such as the correct translation of NEs has proved useful from a linguistic perspective. This is also confirmed by the fact that it helps when combined with other modules. In addition, this module can be used in a pre-processing step in order to normalize expressions (e.g. Time Expressions, NEs). This way, expressions showing a different surface realisation but referring to the same entity/time could be normalized before applying any lexical or syntactic tools and/or resources. This will

lead to a better performance of the NLP tools used and a better performance of VERTa itself.

**Modules Combination**. Experiments have also proved that the most effective combination of modules implies the Lexical Module, the Dependency Module and, in a smaller degree, the N-gram Module and the Semantic Module. Regarding the linguistic information available, the most relevant features to evaluate the adequacy of a segment are features related to lexical semantics and dependency relations included in the Lexical and Dependency Modules, corroborating our initial hypotheses that the interaction between lexical semantics and the dependency relations should account for the meaning of a sentence (see section 1.1). Features related to lexical semantics are unquestionable since they deal directly with meaning, however, although the use of dependency relations might be initially related to syntax, and thus closer to the fluency of a segment, the fact of allowing for flexible dependency matches leads to a broader coverage of different syntactic constructions conveying the same meaning. In addition, both the N-gram and Semantic Modules also contribute, although in a smaller way, to evaluate the adequacy of a segment. In this sense, our hypothesis that other semantic information regarding NE linking, Time expressions identification and normalisation, as well as sentiment analysis, might be suitable to evaluate adequacy, has also been corroborated, even though it is useful in combination with other information (i.e. lexical semantics and dependency relations).

Once experiments on adequacy have been performed and results discussed, the next chapter covers experiments on fluency.

# Chapter 7. Experiments on Fluency

There are MT metrics that assess MT quality in general, others evaluate adequacy and others are more fluency-oriented such as Quirk (2004), BLEU, STM and HWCM (Liu and Gildea 2005), Owczarzak et al. (2007a/b), among others. Once VERTa proved to work well when assessing adequacy, our interest was to analyse if the features included in VERTa could also be relevant to evaluate fluency, that is to say, the degree the output sentence is well-formed according to the rules of the target language (see section 2.1.2). There were some hypotheses as regards the importance of the modules in VERTa that should be used; in addition, checking the usefulness and appropriateness of the internal features constituting those modules was also of our interest. Therefore, the experiments reported in the present chapter were also conducted at segment level.

As mentioned in the paragraph above, experiments are mainly aimed at testing those linguistic features in VERTa that serve best to evaluate fluency. However, one of the experiments performed analyses the impact of using a Language Model (LM) which, despite not being a linguistic feature itself, has been used in some MT metrics to evaluate fluency (Gamon et al. 2005) and whose use is widely extended when references are not available. In addition, in the linguistic analysis (see section 4.2.4) we also found evidence that sometimes segments unfairly get a low score because, despite being grammatically correct, they are too different from the reference translations. Thus, we were interested in exploring the use of Language Models to test how well they correlated with human judgements and if they could be combined with linguistic information.

In the following, all the experiments performed are described and organised as follows: section 7.1 details the data used to conduct these experiments; section 7.2 is devoted to the Lexical Module; section 7.3 explores the Morphological Module; section 7.4 analyses the Dependency Module; section 7.5 covers the N-gram Module; section 7.6 deals with the Semantic Module; section 7.7 is aimed at exploring the use of an LM; section 7.8 studies the combination of the modules in VERTa; and finally, section 7.9 discusses the findings of these experiments.

### 7.1 Data

In order to carry out these experiments, data containing human judgements on fluency was used (refer to section 3.2.1.2 for further details). This data was granted by NIST and LDC, from their NIST 2005 Open Machine Translation (OpenMT) Evaluation campaign. This data includes MT output from Arabic into English from 6 different systems, 4 reference translations and human judgements on fluency. From this data, 100 segments/system were used as development data in order to conduct experiments on fluency, the rest of data was kept to conduct a meta-evaluation of the metric.

### 7.2 Lexical Module

According to the experiments in Chapter 6, the Lexical Module is effective to evaluate adequacy (see section 6.7). This section focuses on the Lexical Module and its portability to the assessment of fluency. First, the linguistic features available in this module are analysed (7.2.1), as well as their importance to evaluate fluency in terms of weights (7.2.2). Finally, a summary of this section is provided (7.2.3)

#### 7.2.1 Linguistic Features in the Lexical Module

The Lexical Module uses mainly the following linguistic information: word-form, lemma, synonyms, partial lemma, and hypernyms/hyponyms whenever necessary. We followed the same process as we did for adequacy (section 6.2.1): the first linguistic feature used is the word-form and the rest of linguistic features are added to this one. First, all features were given the same weight so as to check if VERTa's behaviour was similar to the one shown when evaluating adequacy. Table 48 shows results obtained by the Lexical Module using all the references available.

| Linguistic Features | Pearson Correlation |
|---|---|
| Word_Only | 0.1663 |
| Word + Lemma | 0.1827 |
| Word + Lemma +Synonymy | **0.16433** |
| Word + Lemma + Synonymy + Partial lemma | 0.1751 |
| Word+Lemma+Partial lemma | **0.1851** |

**Table 48** Influence of linguistic features in a 4-reference scenario

It must be highlighted that the results shown in Table 48 are far from being similar to those obtained by the same features when dealing with adequacy, where the correlation of the Lexical Module with human judgments reached 0.7418. This may be due to a) the use of a different corpus with different human judgements, and b) the linguistic knowledge at lexical level is not crucial when assessing fluency since this level focuses mainly on lexical semantics; therefore, the weight assigned to the whole Lexical Module should be rather low.

Regarding the linguistic features used, those that are the most appropriate for this kind of evaluation are the word-form, the lemma of the word and the partial lemma (0.1851). It is noticeable the negative influence of the use of synonymy (in red), which lowers the correlation of the metric with human judgments and, actually, gets the worst results in the 4-reference scenario (see Table 48). These results show a striking contrast to those obtained by this linguistic feature in terms of adequacy, where the use of such a semantic relation had a positive impact on the performance of the metric. In order to check whether these negative scores were consistent across references, each reference was used separately and, as expected, similar results were obtained (see Table 49).

| **Linguistic Features** | **Pearson Correlation** | | | |
|---|---|---|---|---|
| | **Ref. 1** | **Ref. 2** | **Ref. 3** | **Ref. 4** |
| Word_Only | 0.1098 | 0.0674 | 0.0841 | 0.0808 |
| Word + Lemma | **0.1457** | **0.0768** | 0.0912 | 0.0932 |
| Word + Lemma +Synonymy | **0.1281** | **0.0700** | **0.0751** | **0.0894** |
| Word + Lemma + Synonymy + Partial lemma | 0.1302 | 0.0752 | 0.0927 | **0.1015** |
| Word + Lemma + Partial lemma | 0.1379 | 0.0751 | **0.0997** | 0.0980 |

**Table 49** Influence of linguistic features in a single-reference scenario

In order to explore and understand this negative outcome, a thorough analysis of the data was conducted which revealed that although the scores obtained correlated worst with human judgements, from a linguistic point of view, the use of synonyms is absolutely justified, as shown in the example below.

Example 102

HYP: *Rabat* ***refuses attributed*** *to the terrorist acts in the* ***Netherlands***

REF: *Rabat* ***Rejects*** *Acts of Terror in* ***Holland*** *Being* ***Attributed*** *to it*

In 102, there are 2 possible matches as regards synonymy: *refuses – rejects* and *Netherlands – Holland*, which are absolutely correct and do not represent any obstacle in terms of fluency. Matching these synonyms results in the Lexical Module assigning a higher score to the segment, since more lexical items are covered, but it also worsens the correlation with human judgements. The reason for this worsening in the correlation is that the grammaticality of the sentence is not dealt with, mainly because it has nothing to do with lexical semantics, but with other features such as word order, indicating that the problem is not related to the use of synonyms but to the use of the Lexical Module in isolation to evaluate the fluency of a segment. The grammaticality issue in this sentence is mainly related to the position that the Od *the terrorists acts in the Netherlands* occupies after the preposition *to*, instead of being located after the verb *refuse*, as well as the omission of the verb *being* and the pronoun *it*, as illustrated by the edited sentence below.

HYP: *Rabat* ***refuses*** *the terrorist acts in the* ***Netherlands*** ***being*** ***attributed*** *to it*

REF: *Rabat* ***Rejects*** *Acts of Terror in* ***Holland*** *Being* ***Attributed*** *to it*

After some reordering and re-establishing those elements missing, the hypothesis segment is now an excellent candidate both in terms of fluency and adequacy, and the use of synonymy is not a drawback, but a positive feature.

Disregarding the use of synonyms would lead to the omission of certain matches which, from a linguistic point of view, would be totally unfair. Actually, some of those negative synonymy matches might be solved by using other modules, such as the N-gram Module. Likewise, we cannot forget that the rest of modules in VERTa are based on the matches established in the Lexical Module, thus, ignoring information at this level, just on the basis of correlations with human judgements, may lead to disregarding correct matches in other modules. Therefore, although the correlation with human judgements indicates that the use of synonymy is a drawback, we have decided to keep it for the sake of linguistic coherence and good performance of the metric.

Considering those results obtained in both 4-reference and single-reference scenarios, discussion can also be held on the use of partial lemma, which slightly improves the performance of VERTa in a 4-reference scenario, but worsens its performance when using reference 1 and reference 2, individually (see Table 49). A detailed study of the data indicates that similar to the synonyms influence, most of the issues caused by the use of partial lemma can be solved by applying the N-gram Module. On the other hand, the omission of such a feature would be misleading, mainly because strings that are grammatically correct would not be taken into account, as shown in Example 103 with regard to *establish* / *establishment*.

Example 103

HYP: *...negotiations to* ***establish*** *a distinctive partnership between Turkey and the European Union...*

REF: *...negotiations for the* ***establishment*** *of a special partnership between Turkey and the European Union...*

Following the same steps taken when the Lexical Module was developed, the use of hypernyms and hyponyms has also been tested. As illustrated in Table 50, the use of these linguistic features in a 4-reference scenario results in a slight improvement of the correlation of this Lexical Module with human judgements.

| Linguistic Features | Pearson Correlation |
|---|---|
| Word_Only | 0.1663 |
| Word + Lemma | 0.1827 |
| Word + Lemma +Synonymy | 0.1654 |
| Word + Lemma + Synonymy + Partial lemma | 0.1765 |
| Word + Lemma + Synonymy + Partial lemma + Hyps. | **0.1836** |

**Table 50** Influence of hypernyms and hyponyms in a 4-reference scenario

The example below, extracted from the data analysed, proves that the use of such semantic relations helps to slightly widen the coverage of the metric at lexical level.

Example 104

HYP: ...*if the* ***offender*** *is not occupied...*

REF: ...*if the* ***perpetrator*** *is not the occupier...*

In the previous example, the metric considers *perpetrator* as a hyponym of *offender*. Only by using this linguistic feature is the metric capable of considering such a semantic relation.

Once all results are analysed, it can be concluded that all linguistic features contained in the Lexical Module (i.e. word-form, synonyms, lemma, partial lemma, hypernyms and hyponyms) should be considered when assessing fluency, even if their use contradicts the correlation obtained when comparing the metric score with human judgements.

### 7.2.2 Use of Weights

When the Lexical Module was developed, some experiments were carried out in order to check whether different weights should be given to each linguistic feature. Correlations seem to indicate that different weights should be assigned to each type of match, although as reported in Chapter 6, they cannot be considered final weights due to the size of the development corpus; a larger amount of data would be necessary to obtain final weights.

Before adapting the Lexical Module to the evaluation of fluency, our hypothesis was that no different weights were required to perform this kind of evaluation. The reason to formulate such a hypothesis was mainly that fluency considers only the grammaticality of the sentence; in other words, fluency is used to check if the sentence reads good English, disregarding semantics. Therefore the semantic value of the linguistic features used in this module does not seem to affect the fluency of a sentence. However, in order to confirm our hypothesis, we conducted a couple of experiments. First, fluency was evaluated using the same weights assigned when evaluating adequacy, and later the same experiment was carried out, although this time all features were assigned the same weight (see Table 51).

| Linguistic Features | Pearson Correlation |
|---|---|
| Lexical Module using different weights | 0.1781 |
| Lexical Module using the same weights | **0.1836** |

**Table 51** Setting of weights for the Lexical Module when assessing fluency

As shown in Table 51, results support our hypothesis: the version of the Lexical Module that best correlates with human judgements is the one with equal weights for each linguistic feature.

### 7.2.3 Summing Up

This section has explored the use of the Lexical Module to evaluate fluency. The linguistic features used are the same as those used to evaluate adequacy: word-form, lemma, synonyms, hypernyms/hyponyms and partial lemma.

In addition, the importance of each linguistic feature has also been assessed. Whereas in the experiments conducted to evaluate adequacy those features used were assigned different weights according to their semantic value, in the experiments performed to evaluate fluency those features were assigned the same weight each, mainly because their semantic value was not the aim of the experiments. The aim of an evaluation based on fluency is the grammaticality of the segments under study, and their semantic value is of little importance, as confirmed by the experiments performed (see Table 51).

After exploring the Lexical Module, next section studies the Morphological one.

## 7.3 Morphological Module

As reported in section 6.3, the Morphological Module does not correlate well with human judgements on adequacy. In fact, one of our hypotheses was that PoS information in combination with other features might be suitable to evaluate the grammaticality of a segment, thus its fluency. In VERTa, this module should help to restrict the broad coverage of the Lexical Module: PoS tags contain information regarding the lexical category and morphosyntactic features of words, which in English means information regarding tense, number, degree, aspect, mood and voice. This linguistic information can help in terms of fluency to identify issues related to agreement, as well as missing words.

The following explores the linguistic features in this module (7.3.1) and provides a summary of this section (7.3.2).

### 7.3.1 Linguistic Features

The linguistic features used to assess fluency are the same as those used to assess adequacy: a combination of lexical traits with PoS information (see Table 52). The weights used for each combination follow the parameters established for fluency in the Lexical Module, thus all matches receive the same weight.

| Weight | Linguistic Features | Pearson Cor. |
|---|---|---|
| 1 | Word_Only, PoS | 0.2054 |
| 1 | Word + Lemma + Synonymy, PoS | 0.2113 |
| 1 | Word + Lemma + Syn + Hypernymy/Hyponymy, PoS | **0.2173** |

**Table 52** Influence of each type of match in a 4-reference-scenario

As shown in Table 52, the best combination when 4 references are available is the use of all features together with their PoS. In this sense, hypernymy/hyponymy relations also prove effective.

As expected, the correlation with human judgements on fluency is better when the Morphological Module is used, although results are not too distant (see Table 53).

| Module Used | Pearson Correlation |
|---|---|
| Lexical Module | 0.1836 |
| Morphological Module | **0.2173** |

**Table 53** Comparison of the correlation of the metric with human judgements using the Lexical Module and the Morphological Module separately

Next, a couple of examples found in the data analysed supporting the positive effect of the use of the Morphological Module to assess fluency are reported:

a) Identification of subject-verb disagreement as illustrated in the example below.

Example 105

HYP: *...given that the legislative elections **is** the first step...*

REF: *...given that the legislative elections **are** the first step...*

*Legislative elections*, a plural subject should agree with verb *to be* in third person plural, however, in the hypothesis segment the verb form used is that corresponding to the third person singular; thus agreeing with the subject complement instead of the verb. Although meaning might not be affected by such an issue, the fluency of the hypothesis segment is.

b) Identification of missing words as shown in the following example.

Example 106

HYP: ***Abbas ø** the constitutional oath tomorrow, Wednesday*

REF: ***Abbas takes** constitutional oath tomorrow Wednesday*

In the hypothesis segment, the verb of the sentence is missing, thus forcing the PoS tagger to choose another word to work as a verb. In this case, the word *Abbas* in the hypothesis segment has been tagged as a verb, whereas in the reference segment it is a proper noun. As the PoS of both words is different, the Morphological Module has not identified them as a match, and the score obtained by this sentence when only the Lexical Module was used has worsened, reflecting a problem in fluency, which in fact affects adequacy, too.

Although according to the results obtained, the Morphological Module helps in assessing fluency, higher results were expected. The data analysed revealed that some errors in the automatic PoS tagging were affecting negatively the performance of this module. These errors were mainly related to the tagging of headings, which in the reference translations appear in capital letters, whereas in the hypothesis segments tend to appear in lower case, as shown in Example 107.

Example 107

HYP: *Ardogan-NNP confirms-VBZ that-IN Turkey-NNP **will-MD reject-VB** any-DT pressures-NNP to-TO urge-VB them-PRP to-TO **recognize-VB** the-DT Cyprus-NNP*

REF: *Erdogan-NNP Confirms-VBZ That-IN Turkey-NNP* ***Will-NNP Reject-NNP*** *Any-DT Pressures-NNP To-TO Encourage-VB It-PRP To-TO* ***Recognize-NNP*** *Cyprus-NNP*

Although the hypothesis and reference segment are very close, and the hypothesis segment could have achieved a high score as regards fluency, the fact that the PoS tagger had problems to handle *Will*, *Reject* and *Recognize* in the reference sentence prevents those three positive matches. Consequently, the score obtained by the hypothesis sentence is lower than it should actually be.

### 7.3.2 Summing Up

Experiments have shown that the all linguistic features in this module are suitable to evaluate the fluency of a segment: word-form, lemma, synonyms, hypernyms and hyponyms. In addition, all features should be assigned the same weight.

To conclude, the Morphological Module correlates better with fluency judgements than the Lexical Module, although there is no big difference.

Once the Morphological Module has been explored, the next section is devoted to the Dependency Module.

## 7.4 Dependency Module

There are several researchers that have proposed using dependency relations to evaluate the fluency of a segment, such as Owczarzak et al. (2007a/b). However, in our study the Dependency Module has proved effective to assess adequacy, so the most natural next step was checking whether it was also useful to assess fluency and whether the same matches, rules and parameters had to be used. Experiments checked first the appropriateness of the different types of matches (7.4.1); then, the importance of dependency tags (7.4.2); and later, the suitability of the extra-rules added to this module (7.4.3). Finally, a summary of this section is provided (7.4.4).

### 7.4.1 Dependency Matches

The first experiment was aimed at checking the suitability of the different types of matches available in a 4-reference scenario. To conduct this experiment, both dependency labels and types of matches were assigned the same weight. The results obtained were correlated with the human judgements on fluency to check which type(s) of match(es) performed best (see Table 54).

| Type of Match | Pearson Correlation |
|---|---|
| Exact match | 0.3108 |
| Exact + No_label | 0.3038 |
| Exact + No_label + No_mod | **0.3218** |
| Exact + No_label + No_mod + No_head | 0.2930 |
| Exact + No_mod | **0.3228** |

**Table 54** Influence of each type of match in the Dependency Module in a 4-reference scenario

As shown in Table 54, although there is no remarkable difference, the combinations that correlate best with human judgements are first the combination **Exact match + No_mod match**, followed closely by the combination **Exact match + No_label match + No_mod match**. It must be highlighted that the addition of the No_label match and, especially, the No_head match worsens the performance of the metric, even dropping it to a lower score than that obtained when only the Exact match is being used; whereas the use of the No_mod match influences positively and increases the correlation from 0.3038 to 0.3218. So as to confirm those results obtained, a single-reference scenario was used (see Table 55).

| Match Types | Pearson Correlation | | | |
|---|---|---|---|---|
| | **Ref. 1** | **Ref. 2** | **Ref. 3** | **Ref. 4** |
| Exact match | 0.2265 | 0.1613 | 0.2097 | 0.2159 |
| Exact + No_label | 0.2262 | 0.1485 | 0.2091 | 0.2158 |
| Exact + No_label + No_mod | **0.2527** | **0.1968** | **0.2336** | **0.2594** |
| Exact + No_label + No_mod + No_head | 0.2260 | 0.1377 | 0.1978 | 0.2135 |
| Exact + No_mod | **0.2539** | **0.2022** | **0.2299** | **0.2606** |

**Table 55** Influence of each type of match in the Dependency Module in a single-reference scenario

As expected, similar results were obtained in a single-reference scenario where the combinations **Exact match + No_mod match** and **Exact match + No_label match + No_mod match** were those achieving the best results.

It is noticeable the different types of matches used when assessing adequacy and when assessing fluency. When evaluating adequacy (see section 6.4.1), correlations indicated that the No_mod match slightly worsened correlation scores whereas the No_label match, and especially the No_head match strongly improved them. Finally, linguistic analysis confirmed the usefulness of all matches when evaluating adequacy. As regards fluency, a linguistic analysis has also been carried out in order to finally choose the best combination of matches. Next, every type of match is analysed in detail.

#### 7.4.1.1 No_label Match

When the Dependency Module was used to assess adequacy, the No_label match was extremely helpful because it allowed for identifying different syntactic structures which expressed the same meaning. However, according to correlations, it does not seem to have the same effect on the fluency of the segment. The reason might be that when evaluating adequacy, even if those structures were not totally grammatical, they were able to convey the same meaning, thus improving the correlation with human judgements on adequacy. On the contrary, using this match could have a negative effect when dealing with fluency, as illustrated in Example 108.

Example 108

HYP: *Sharp controversy in Morocco on press reports about* ***the incomes of ⌀ King***

REF: *Sharp debate in Morocco on Press Reports on* ***King's Income***

The NP *the incomes of King* in the hypothesis segment differs syntactically from the NP *King's income* in that the former contains a PP working as a complement of *incomes* to express possession, whereas the latter contains a NP genitive premodifying it. Both structures are comparable from a semantic point of view; however, the NP *the incomes of King* is not a fluent chunk since a determiner should precede the noun *King*. Therefore, the No_label match links two triples identical in meaning but which imply a drawback at fluency level (see Table 56).

| Hypothesis | Reference | Match |
|---|---|---|
| prep_of(income,King) | poss(income,King) | No_label match |

**Table 56** Dependency triples match corresponding to Example 108

A similar example but involving the omission of a preposition is illustrated by Example 109. In the hypothesis sentence, the preposition *on* introducing the adjunct of time is missing, thus affecting fluency and at some point also adequacy.

Example 109

HYP: *14 people were killed* ***⌀ Sunday****...*

REF: ***On Sunday****, 14 people were also killed...*

As shown in Table 57, the corresponding triples can be matched if the No_label match is used, resulting in a fluency issue.

| Hypothesis | Reference | Match |
|---|---|---|
| tmod(killed,Sunday) | prep_on(killed,Sunday) | No_label match |

**Table 57** Dependency triples match corresponding to Example 109

The example above affected determiners and prepositions but other elements can be affected, such as the omission of the subordinator *that* to introduce a clause

complement, when it should not be omitted. Such is the case of Example 110 where the subordinator in the hypothesis sentence has been omitted resulting in an ungrammatical sentence.

Example 110

HYP: *The Iranian Foreign Ministry spokesman* ***stated*** *yesterday, Sunday,* ***ø Iranian Foreign Minister Kamal Kharazi said****...*

REF: *A spokesman for the Iranian Foreign Minister* ***stated*** *yesterday, Sunday,* ***that Foreign Minister Kamal Kharazi****...*

Although from a semantic point of view, the use of the No_label match may help to match the dependency triples corresponding to those structures, as regards fluency, this match has a negative effect.

#### 7.4.1.2 No_mod Match

In this section the positive and negative effects of using the No_mod match are analysed.

Regarding the positive effects of this match, it mainly widens the coverage of matches due to lexical semantics. In other words, this match establishes relations between words that were disregarded in the Lexical Module but which are semantically related (similar to its performance in section 6.4.1.2). Example 111 shows this widening of the lexical coverage.

Example 111

HYP: *He confirmed that "the terrorists Moroccans had also* ***relations*** *with networks...*

REF: *He said: "Moroccan terrorists also have* ***links*** *with foreign networks...*

The words *relations* and *links* are semantically related but are not covered in WordNet; therefore the Lexical Module does not consider them as a possible similarity. However, as reported in Table 58, the No_mod match allows for this similarity.

| Hypothesis | Reference | Match |
|---|---|---|
| dobj(had,relations) | dobj(have,links) | No_mod match |

**Table 58** Dependency triples match corresponding to Example 111

All in all, the widening of lexical coverage that this match allows is more related to adequacy than to fluency. Linguistic analysis has revealed that using this type of match to evaluate fluency might bring more negative effects than positive ones, as reported below:

a) Word order at clause level. Some MT engines have problems when reordering constituents inside the sentence. In Arabic, the subject follows the verb, whereas in English the canonical position of the subject in a clause is before the verb. A failure in the reordering of the subject and verb positions in the sentence leads to both fluency and adequacy issues, as exemplified below.

Example 112

HYP: *Lusaka 8 December / Xinhua /* ***celebrated the Common Market for Eastern and Southern Africa (COMESA)*** *[...] anniversary*

REF: *Lusaka December 8 / Xinhua /* ***The Common Market of Eastern and Southern Africa (COMESA) [...] celebrated*** *its anniversary*

In the hypothesis segment, the subject *the Common Market for Eatern and Southern Africa (COMESA)* appears in the position of the object, following the verb, which violates the position of the subject in the canonical English sentence. This failure in word order affects the fluency of the sentence because it does not follow the English word order. The No_mod match allows for the similarity between the triples in Table 59, establishing a misleading similarity match.

| Hypothesis | Reference | Match |
|---|---|---|
| NO MATCH | nsubj(celebrated,Market) | No match |
| dobj(celebrated,Market) | dobj(celebrated,anniversary) | No_mod match |

**Table 59** Dependency triples match corresponding to Example 112

b) Pronoun mistake. MT engines may have problems to translate a pronoun in its correct form. An example is the confusion between possessive adjectives and object pronouns. Example 113 shows that the MT engine failed in translating the object pronoun correctly, instead of using the object pronoun *us*, it used the possessive adjective *our*.

Example 113

HYP: *...the European Union cannot address **our** across new conditionalities...*

REF: *...the European Union cannot address **us** by imposing new conditions...*

| Hypothesis | Reference | Match |
|---|---|---|
| dobj(address,our) | dobj(address,us) | No_mod match |

**Table 60** Dependency triples match corresponding to Example 113

The No_mod match, as reported in Table 60, allows for a false similarity match, which will have a negative effect in the grammaticality judgement of the sentence, thus affecting negatively in the fluency score. It must be highlighted as well, that by disregarding this kind of match we deal with a problem that cannot be tackled by the Morphological Module, since the PoS tags used do not distinguish between different types of pronouns.

c) Omission of determiners, prepositions and conjunction. It is common that conjunctions, prepositions and determiners are not translated or are mistranslated. The comparison between the hypothesis and reference segments in Example 114 shows these untranslated items.

Example 114

HYP: *...will discuss the agenda for the meeting ø a draft ø ø final statement.*

REF: *...will discuss the meeting's agenda **and** a draft **of the** final statement.*

The MT system fails in translating the coordinating conjunction *and*, which leads to the No Match in Table 61. In addition, the preposition *of* and the determiner *the* have also been disregarded, which results in a different dependency analysis. In the hypothesis segment, the head of the NP is *statement* whereas *draft* and *final* are just premodifiers.

In the reference segment, *draft* is the head of the NP and the chunk *of the final statement* is a PP working as a complement of *draft*. When comparing the triples provided by the dependency parser, these different analyses can be observed in the No_mod match, where de determiner *a* in the hypothesis segment relates to the word *statement* instead of *draft*. Likewise, the word *draft* as a modifier of *statement* cannot be matched to any reference triple, highlighting this mismatch of structures. The omission of such lexical items clearly affects the fluency of the sentence.

| Hypothesis | Reference | Match |
|---|---|---|
| NO MATCH | conj_and(agenda,draft) | No match |
| **det(statement,a)** | **det(statement,the)** | **No_mod match** |
| **amod(statement,draft)** | **NO MATCH** | **No match** |
| amod(statement,final) | amod(statement,final) | Exact match |
| NO MATCH | prep_of(draft,statement) | No match |
| dep(meeting,statement) | NO MATCH | No match |

**Table 61** Dependency triples match corresponding to Example 114

Once data has been analysed, the only positive effect identified is a broader coverage of lexical matches. It seems that even if this match slightly improves the correlation with human judgements, its effects from a linguistic point of view are rather opposite, since it may lead to misleading matches regarding the grammaticality of a sentence.

#### 7.4.1.3 No_head Match

The last type of match considered is the No_head match, which was already tested when dealing with adequacy. This type of match implies that the triples compared share the same label and modifier, but their head differs. According to the results obtained when the correlation with human judgements was calculated, the use of this match does not seem to help; however, data has been analysed from a linguistic point of view to check whether there were some positive effects and which were the negative ones.

On the one hand, the positive effect of using this type of match is mainly widening the coverage at lexical level, similar to the No_mod match. The Dependency Module relies on lexical semantic relations available in WordNet; however some semantic relations

such as near synonymy are not covered and the No_head match allows for their coverage, as shown in Example 115.

Example 115

HYP: *We will discuss **this file** during the accession negotiations.*

REF: *We will discuss **this dossier** in the course of membership negotiations.*

| Hypothesis | Reference | Match |
|---|---|---|
| nsubj(discuss,we) | nsubj(discuss,we) | Exact match |
| aux(discuss,will) | aux(discuss,will) | Exact match |
| **det(file,this)** | **det(dossier,this)** | **No_head match** |
| **dobj(discuss,file)** | **dobj(discuss,dossier)** | **No_mod match** |

**Table 62** Dependency triples match corresponding to Example 115

The semantic relation between *file* and *dossier* cannot be established by using WordNet in the Lexical Module. Only a more flexible triple match (i.e. No_head match and No_mod match) can help in establishing such a relation, as shown in Table 62 where *file* and *dossier* are related thanks to the No_head and No_mod matches. Unfortunately, in the data analysed this has been the only positive influence of using the No_head match found; actually, most of the evidence analysed shows a negative effect, which confirms the drop in the correlation with human judgements.

On a different note, recognizing and extracting No_head matches when evaluating fluency may positively help to identify grammatical issues:

a) The No_head match can be a clue to identify parts of the sentence that have not been translated and therefore, form ungrammatical chunks. A clear example is the omission of the verb due to the failure of the MT engine to translate it, as shown in Example 116.

Example 116

HYP: *...The Turkish government Tayeb Ardogan **ø today** Wednesday that...*

REF: *...Turkish Prime Minister Recep Tayyip Erdogan **announced today**, Wednesday, that...*

| Hypothesis | Reference | Match |
|---|---|---|
| tmod(Tayeb,today) | tmod(announced,today) | No_head match |

**Table 63** Dependency triples match corresponding to Example 116

As reported in Table 63, the fact that the verb *announced* is missing in the hypothesis segment makes the dependency parser consider *Tayeb* as a verb and *today* as a time modifier of *Tayeb*, whereas in the reference segment, containing the verb *announced*, the dependency parser correctly analyses the word *today* as a time modifier depending on the verb *announced*. Therefore, the No_head match indicates that there is some kind of grammatical issue in the hypothesis segment.

Likewise, this type of match may also hint the omission of the subject in the hypothesis segment, as illustrated in Example 117.

Example 117

HYP: *...in reference to which resulted in* ***the May 16, 2003 ø killed*** *45 people, including* ***12 suicide ø*** *in the White House.*

REF: *...alluding to* ***the attacks of May 16, 2003 attacks which killed 45 people*** *among them* ***12 suicide bombers*** *in Casablanca.*

| Hypothesis | Reference | Match |
|---|---|---|
| det(people,the) | det(attacks,the) | No_head match |
| nn(people,May) | NO MATCH | No match |
| num(people,16) | num(May,16) | No_head match |
| prep_in(resulted,people) | NO MATCH | No match |
| num(people,2003) | num(attacks,2003) | No_head match |
| amod(people,killed) | NO MATCH | No match |
| num(people,45) | num(people,45) | Exact match |

**Table 64** Dependency triples match corresponding to the hypothesis segment in Example 117

In the hypothesis segment, the subject before the verb *killed* has not been translated, which makes the dependency parser analyse the chunk *the May 16, 2003 killed 45 people* as a large NP, resulting in the high number of No_head matches and No matches

(Table 64). In the reference segment, where the verb is present, the resulting triples corresponding to the subject and object dependency relations get a No match when compared to the hypothesis triples (Table 65).

| Reference | Hypothesis | Match |
|---|---|---|
| det(attacks,the) | det(people,the) | No_head match |
| prep_to(alluding,attacks) | NO MATCH | No match |
| prep_of(attacks,May) | NO MATCH | No match |
| num(May,16) | num(people,16) | No_head match |
| num(attacks,2003) | num(people,2003) | No_head match |
| **nsubj(killed,attacks)** | **NO MATCH** | **No match** |
| rcmod(attacks,killed) | NO MATCH | No match |
| num(people,45) | num(people,45) | Exact match |
| **dobj(killed,people)** | **NO MATCH** | **No match** |

**Table 65** Dependency triples matches corresponding to the reference segment in Example 117

Not only does the No_head match indicate the omission of an element at a clause level, but it may also identify the omission of an element at a phrase level. That is the case of the noun *bombers* (in red) in the same Example 117. In the hypothesis segment, this word has been omitted resulting into an ungrammatical NP. Although from a semantic point of view this is not a big issue since the meaning of this NP is not affected much, its grammaticality is. Table 66 shows that the quantifier *12* is related to the noun suicide in the hypothesis segment and *bombers* in the reference segment, respectively, resulting into a No_head match. In addition, the triple stating the dependency relation between the noun *suicide* and the noun *bombers* cannot match any triple in the hypothesis segment.

| Hypothesis | Reference | Match |
|---|---|---|
| num(suicide,12) | num(bombers,12) | No_head match |
| NO MATCH | nn(bombers,suicide) | No match |

**Table 66** Dependency triples match corresponding to the NP *12 suicide bombers* in Example 117

As stated above, the meaning of this part of the sentence is not affected, but the fact that the singular noun *suicide* is preceded by number *12* results in the lack of agreement between the numeral and the noun, which affects the fluency of the hypothesis segment.

b) The No_head match can be a hint of a problem in word order. Example 118 illustrates wrong word order at phrase level.

Example 118

HYP: ...*suspected of* ***some detainees Moroccans***...

REF: ...***some Moroccan detainees*** *are clearly suspected of*...

In the hypothesis translation there is a problem of word order in the NP *some detainees Moroccans*, since *Moroccans* follows the noun *detainees* instead of preceding it. In addition, the plural suffix –s has been added to the word *Moroccan*. The triples corresponding to this NP result into a No_head match and a No match (see Table 67).

| Hypothesis | Reference | Match |
|---|---|---|
| det(Moroccans,some) | det(detainees,some) | No_mod match |
| amod(Moroccans,detainees) | NO MATCH | No match |
| NO MATCH | amod(detainees,Moroccan) | No match |

**Table 67** Dependency triples match corresponding to Example 118

The meaning of this NP is not affected by this wrong order, but its fluency is.

Once the linguistic analysis has been performed, we must question the use of matches other than the Exact match, even if this linguistic approach contradicts correlation scores. As shown in sections 7.4.1.2 and 7.4.1.3, more flexible matches – especially the No_head and No_mod matches – only allow for a broader coverage of lexical semantic relations but they may lead to incorrect matches as regards the fluency of a segment. That is why the use of the Exact match in isolation is preferred in order to evaluate fluency, even if it might be too restrictive in some cases. Contrary to the use of the Dependency Module to assess adequacy, where the different types of matches help to maximize the coverage of the meaning of the sentence by allowing for different syntactic structures conveying the same meaning, in the case of fluency, the more

restrictive the type of match is, the better the results obtained, at least according to the linguistic analysis.

After analysing the different types of matches in detail, next the dependency labels used are studied.

### 7.4.2 Dependency Labels

When the Dependency Module was first developed to assess adequacy, dependency labels were grouped in three sets and different weights were assigned, following He et al. (2010)'s approach. However, the results obtained in our experiments showed that grouping dependency labels and assigning different weights to each group did not correlate well with human judgements. In fact, the final decision taken was that most of the dependency labels were worth the maximum weight, except for a reduced number of labels, namely *dep*, *det* and _ which were assigned 0.5. These groupings were revisited when the Dependency Module was used to assess fluency and this time the use of three different sets being assigned different weights correlated well with human judgements on fluency. These sets followed the initial groupings, although weights for MID and LOW sets changed to 0.5 and 0, respectively.

- TOP: dependency relations affecting the arguments of the verb, auxiliary verbs (both modal and non-modal), and copular verbs. [nsubj, dobj, aux, ccomp, rcmod, auxpass, nsubjpass, csubjpass, xsubj, cop, advcl, agent, appos, neg, parataxis, csubj, iobj, acomp, expl, attr, purpcl, root]: **1**

- MID: dependency relations affecting adjuncts and phrase level modifiers and complements [amod, nn, prep, prep_*, conj_*, conj, advmod, xcomp, prt, mark, pobj, cc, infmod, rel, pcomp, prepc_*, abbrev, partmod, ref, tmod]: **0.5**

- LOW: dependency relations related to punctuation marks, determiners and unlabeled constituents [dep, det, discourse, punct, complm, poss, num, number, predet, npadvmod, quantmod, possessive, measure, preconj, mwe, _]: **0**

The fact that dependency matches have been restricted, and only the Exact match is being used, results in a stronger influence of the position of the dependency labels inside the syntactic tree. It seems that those grammatical chunks that function as top

dependency relations are more valued by human judges than those that, although being grammatical, correspond to low-level dependency relations; to put it simply, a human judge would penalise the omission of the subject more strongly than the omission of a determiner or the modifier of a noun.

As regards the impact that using the different sets of labels had on the correlation with human judgements on fluency, this underwent a considerable increase. The correlation of the Dependency Module using the Exact match and equal-weight dependency relations was **0.3108**, whereas it increased to **0.3802** when dependency relations were assigned different weights.

### 7.4.3 Rules

As reported in Chapter 5, aimed at the metric description, the Dependency Module allows for the use of some language-dependent rules that help to match different syntactic structures (see section 5.2.3). From a linguistic point of view, these rules are especially important when the Exact match is the only match used because they help to make the Dependency Module a little more flexible. From the rules available those that seem to slightly increase the correlation with human judgements are those referring to possessive constructions, the active-passive alternation and the dative-ditransitive alternation, which help to move from **0.3802** up to **0.3830**.

### 7.4.4 Summing Up

Once experiments have been performed and results analysed, we can conclude that, from a linguistic point of view, the most effective type of match to evaluate the fluency of a segment is the Exact match. Linguistic analysis has shown that the rest of matches are too flexible since they allow for matching structures that are not grammatically correct. In addition, dependency labels should be organized into three categories (i.e. top nodes, middle nodes and ultimate nodes), which receive different weights: 1, 0.5 and 0, respectively. Finally, from the language-dependent rules added, those that allow for comparing different syntactic structures conveying the same meaning slightly improve the correlation with human judgements, since they broaden the restrictive coverage of the Exact match.

So far, the Dependency Module is the one that has been strongly modified, next the N-gram Module and its corresponding features will be explored.

## 7.5 N-gram Module

In section 6.5, aimed at the use of the N-gram Module to assess adequacy, it was highlighted that although the correlation with human judgements obtained was not bad, this module did not achieve the same results as the Lexical Module. This was mainly due to the restrictive nature of the N-gram Module, which prevented some meaningful segments or parts of segments from getting a good score. Some of the data analysed showed that this module was particularly useful to identify wrong word order, as well as missing words, such as prepositions and determiners, which might be useful to check the grammaticality of the sentence. Next, experiments to confirm this hypothesis are conducted.

This section is organised as follows: section 7.5.1 describes the N-gram matches and section 7.5.2 briefly summarizes this section.

### 7.5.1 N-gram Matches

In this section the N-gram Module is tested as regards fluency by means of the same experiments carried out when adequacy was assessed:

a) Computing n-grams over lexical items

b) Computing n-grams over lexical items and PoS combinations

c) Computing n-grams over PoS

Similarly, different n-gram distances are tested in order to decide which correlates best with human judgements on fluency:

a) From 2grams to sentence-length-grams

b) Only 2grams

c) 2grams and 3grams

d) 2grams, 3grams and 4grams

The results of these experiments are reported in Table 68.

| N-grams Length | N-grams on Lexical Items | N-grams on Combination Lexical Items + PoS | N-grams on PoS |
|---|---|---|---|
| **2grams to sentence-length grams** | 0.3112 | 0.3166 | **0.3459** |
| **2grams** | 0.2204 | 0.2174 | 0.1645 |
| **2grams and 3grams** | 0.2181 | 0.2106 | 0.1938 |
| **2grams, 3grams and 4grams** | 0.2191 | 0.2064 | 0.21296 |

**Table 68** Correlation of the N-gram Module with human judgements on fluency

As shown in Table 68, those results that correlate best with human judgements are those obtained when n-grams are calculated over PoS (0.3459). This is due to the close link between the information about morphosyntactic features provided by the PoS tag, word order and the grammaticality of the segment.

Moreover, according to the three experiments, it is unquestionable that the longer the n-gram distance, the better the correlation with human judgements, at least in English, a language with a rather fixed word order. Therefore, the restriction imposed by the N-gram Module proves to be valuable in order to assess the grammaticality of a sentence.

From a more linguistic approach, experiments confirmed that the N-gram Module is extremely useful to control word order, as shown in the example below.

Example 119

HYP: *[He added that Iraq's]* ***neighbours six*** *[will participate in the conference]...*

REF: *[He added that Iraq's]* ***six neighbouring*** *countries [will participate in the Conference]...*

The N-gram Module is able to match those chunks between brackets *He added that Iraq's* and *will participate in the conference*. However, in the hypothesis segment the numeral *six* follows the noun *neighbours* resulting into a disfluent sentence. In addition,

there is no possible n-gram match for the word *countries* in the reference segment since it has not been translated in the hypothesis.

If we compare results obtained by the N-gram Module for fluency and adequacy, we notice that they are rather opposite (see Table 69), which shows that the information needed in both tasks has to be different.

| **N-gram Length** | **N-gram Module to Assess Fluency** | | | **N-gram Module to Assess Adequacy** | | |
|---|---|---|---|---|---|---|
| | **N-grams over lexic.** | **N-grams over lexic. + PoS** | **N-grams over PoS** | **N-grams over lexic.** | **N-grams over lexic. + PoS** | **N-grams over PoS** |
| **2grams to sentence-grams** | 0.3112 | 0.31661 | **0.3459** | 0.4109 | 0.3921 | 0.4187 |
| **2grams** | 0.2204 | 0.2174 | 0.1645 | **0.7019** | 0.6477 | 0.5610 |
| **2grams + 3grams** | 0.2181 | 0.2106 | 0.1938 | 0.6805 | 0.63058 | 0.5909 |
| **2grams, 3grams + 4grams** | 0.2191 | 0.2064 | 0.2129 | 0.6587 | 0.6113 | 0.5993 |

**Table 69** Comparison of N-gram Module assessing fluency and N-gram Module assessing adequacy

Results indicate that linguistic features used must vary depending on the goal of the evaluation. If the meaning of the sentence is assessed, lexical features and shorter n-grams must be favoured, whereas if the grammaticality of the sentence is assessed, PoS features and longer n-grams must be used.

### 7.5.2 Summing Up

Once experiments have been carried out, it can be concluded that the N-gram Module has proved to be the most effective module to evaluate the fluency of a segment, so far. Experiments have also shown that larger n-grams are more appropriate to evaluate the grammaticality of a segment as they account for word order and that n-grams work

better when calculated over PoS tags. Thus, this couple of features must be combined when evaluating fluency.

In the following, the Semantic Module is explored.

### 7.6 Semantic Module

The features aimed at checking sentence semantics are obviously not appropriate to check the sentence fluency as the features taken into account – NEs, Time Expressions and Sentiment analysis – do not have any influence on the grammaticality of the sentence. In fact, the tools used might disregard certain grammatical features, such as the Time Expression tool which is able to capture the similarity between two Time Expressions even if one of them is not grammatically accurate (e.g. the month is written using lower case). Actually, if correlations with human judgements based on fluency are calculated, most of the modules obtain a negative correlation (see Table 70), indicating that they are useless to evaluate fluency.

| Semantic Module | Pearson Correlation |
|---|---|
| NER | **-0.0590** |
| NEL | **0.1568** |
| TIMEX | **-0.0177** |
| Sentiment Analysis | **-0.0430** |
| Combination of Semantic components | **0.0982** |

**Table 70** Correlation between the Semantic Module and human judgements on fluency

The only feature that slightly helps is Named Entity Linking (NEL) (0.1568) which, as explained in sections 3.2.2.1 and 5.2.5, links NEs with Wikipedia pages. Thus, if a NE has not been translated properly (i.e. proper nouns) it will not be found in Wikipedia.

Even if all features are combined the correlation is really low (0.0982), indicating that features related to semantics are clearly not suitable to evaluate fluency.
On the other hand, MT metrics have been using LMs to address fluency issues. In the next section an experiment is carried out on the use of LMs for the task of fluency.

## 7.7 Language Model Module

There are hypothesis segments that are very different from the translation references, but which are still correct. This turns into a problem, particularly in the case of fluency (see section 4.2.4). The same meaning can be expressed in many different ways, all of them grammatical but which might not be covered by the reference translations available. Thus, using an LM instead of reference translations seems particularly useful to tackle this issue. By using an LM we aim at accounting for those segments that, even being syntactically different from their corresponding reference translations, are still fluent.

In this experiment, we do not try to find similarity matches between the hypothesis and reference segments, neither try to compare them. We use an LM to calculate the degree (log probability) to which the hypothesis segment is expected compared to what occurs in the corpus used to build the LM.

For this experiment, we tried three different LMs, all of them based on ngrams over lexical items:
- Europarl LM[41]. This LM was used in the WMT13 quality estimation task as a baseline feature. This resource was built from the Europarl data released as part of WMT11.
- News LM[42]. This LM was also used in the WMT13 quality estimation task as a baseline feature. This resource was built from the news data released as part of WMT11.
- Google N-grams LM[43]. This LM is based on the Google N-grams corpus.

Correlations obtained (see Table 71) show that the LM that correlates best with human judgements on adequacy is the news-based LM, which is not surprising since it is a domain-related LM.

[41] http://www.quest.dcs.shef.ac.uk/quest_files/lm.europarl-nc.en
[42] http://www.quest.dcs.shef.ac.uk/quest_files/de-en/news.3gram.en.lm
[43] https://code.google.com/p/berkeleylm/

| Language Model | Pearson Correlation |
|---|---|
| Europarl LM | 0.1047 |
| News LM | **0.2579** |
| Google N-grams LM | 0.2081 |

**Table 71** Correlation with human judgements using LMs

Although the news-based LM obtains the best correlation (0.2579) from the three LM tested, we expected the correlation to be higher. If compared to the Dependency and N-gram Modules, these still show better correlations (0.3830 and 0.3459, respectively). It is not surprising that our N-gram Module works better than LMs, since the former uses N-grams calculated over PoS, which is more appropriate to control word order when assessing the fluency of a segment.

Thus, we would like to try if the LM Module might help when combined with other modules.

### 7.8 Modules Combination

Once all modules were analysed separately, the next step was exploring how they should interact in order to get the best combination.

As shown in Table 72, those modules that correlate best with human judgements on fluency are the dependency and N-gram Modules. This corroborates our initial hypothesis (see section 1.1) that the linguistic information accounting for the grammatical structure of a sentence and word order should be the most appropriate to assess fluency. First, the Dependency Module accounts for the phrase and sentence structure; and second, the N-gram Module accounts for word order, which is an important characteristic in English. In addition, we must highlight that in this case, the n-grams are calculated over PoS tags, thus, inflectional morphology, syntactic categories and morphosyntax are also taken into account.

| Module | Pearson Correlation |
|---|---|
| Lexical Module | 0.1836 |
| Morphological Module | 0.2173 |
| Dependency Module | **0.3830** |
| N-gram Module | **0.3459** |
| Semantic Module | 0.0982 |
| LM Module (news-based) | 0.2579 |

**Table 72** Pearson correlation on fluency per module

Hence, it was clear that both modules should interact, but in order to justify their combination, a couple of possibilities were considered: in the first experiment, all modules were used and they were assigned the same weight, whereas in the second one, in order to calculate an upper-bound for the weight tuning, all possible weight combinations were tuned automatically using a 0.01 step. Table 73 shows the results of these experiments and that the best combination obtained was rather different from that expected, since finally four modules contributed to the evaluation: the Morphological Module, the Dependency Module, the N-gram Module and the LM Module. Weights assigned to each module are shown in Table 73.

| Modules Combination | Pearson Correlation |
|---|---|
| All modules - maximum weight | 0.4034 |
| Lexical M. (0), Morphological M.(0.04), Dependency M.(0.37), N-gram M.(0.29), Semantic M. (0) and LM M.(0.30) | 0.4341 |

**Table 73** Combination of modules and weights assigned

The Dependency Module is clearly the module that most contributes to the performance of the metric, next is the LM Module followed closely by the N-gram Module. Finally, the Morphological Module contributes slightly to the performance of the metric. Both N-gram Modules and LM Modules show a similar performance, although the first accounts for PoS n-grams while the second focuses on n-grams over lexical items that might not appear in the reference translations. We can say, therefore, that they complement each other. The small contribution of the Morphological Module can also

be explained because a) the N-gram Module is already taking into account PoS information, covering issues such as agreement; and b) as explained above, English does not show a rich inflectional morphology, thus individual PoS matching is not that important.

Some of the grammaticality issues that could be detected with the use of the modules combination reported above are the following:

a) Sentences without subject. In English all sentences must contain a subject in order to be grammatical, however this is still a problem for some machine translation engines which are either unable to translate the subject or provide an incorrect translation, mainly using 3rd person singular pronoun *he* in its place. Missing subjects affect not only adequacy but also fluency, as shown in Example 120. The use of the Dependency Module with the Exact match and higher weight to top-level dependency relations help to detect this type of issues.

Example 120

HYP: *In an interview with the newspaper le "**ø** **confirmed** that the persons involved in terrorist cases in the Netherlands...".*

REF: *In an interview with the "Aujourd'hui le Maroc" newspaper, **Bouzoubaa** **stressed** that the people involved in the terror cases in Holland....*

In this example, *Bouzoubaa*, the subject of the main clause in the hypothesis sentence, is missing, thus affecting the grammaticality of the segment.

b) Lexicogrammatical patterns. The type of complements that verbs take plays an important role in the grammaticality of a sentence. Examples 121 and 122 illustrate their importance.

Example 121

HYP: *He **said** Ardogan station "TV" television that "the European Union cannot address..."*

The default pattern that verb *say* enters is SVObl (say something to somebody), however, this verb can also subcategorize for a clause complement realised by a that-clause. In this case, the pattern would be SVClCompl (say that....). Thus, the dependency parser analyses the chunk *Ardogan station TV television that "the European Union cannot address...* as the direct object of the main verb, where *address* and *television* are linked by the dependency tag *dep* which indicates that this is a weird grammatical structure. In this case, the verb used should have been *tell* which accepts *tell somebody something*. Furthermore, it must also be noticed that *Ardogan* should occupy the subject position instead of *He*.

In Example 122, attention should be paid to the chunk in bold *see each warned of Morroccan terrorist acts commited in the Netherlands*.

Example 122

HYP: *The minister added, "which is why I said to* ***see each warned of Morroccan terrorist acts committed in the Netherlands****."*

The verb *see* subcategorizes for a direct object, however, in the chunk there is no Noun that could work as the head of the direct object, which is due to a bad translation. As a consequence, the analysis provided by the dependency parser links *see* and *warned* by means of the tag *dep*, indicating, again, that there is a weird grammatical structure.

c) Word order of immediate constituents. Sometimes a constituent itself might show a correct internal grammatical structure, but it might occupy an ungrammatical position at clause level resulting in an ungrammatical sentence. Example 123 illustrates this fact.

Example 123

HYP: *Baghdad 24 – 12 (AFP) –* ***accused*** *[Shiite leader of the hardline young issued] [today, Friday,] [Israel and the United States and Britain] [of being behind the bloody attacks against the cities, Najaf and Kerbala last Sunday, which claimed the lives of 66 people dead and some 200 injured].*

REF: *Baghdad 12-24 (AFP) – [The young radical Shiite leader Muqtada Al-Sadr]* ***accused*** *[today, Friday], [Irael, the United States and Britain] [of being behind the*

*bloody attacks that targeted the two cities of Najaf and Karbala last Sunday and in which 66 people were killed and about 200 injured].*

In Example 123, the NP realising the subject has not been translated properly and, in addition reordering is needed, as it occupies the position of the object. Consequently, the sentence is clearly disfluent and although some of the immediate constituents present a correct internal grammatical structure, the grammaticality of the whole sentence is clearly affected. The grammaticality of the constituents internal structure is mainly captured by the N-gram Module, which provides better results (see Table 74) than the Dependency Module which is clearly affected by the ungrammatical position of the immediate constituents.

| Modules | Score Obtained |
|---|---|
| N-gram Module | 0.2702 |
| Dependency Module | 0.1690 |

**Table 74** Score per module corresponding to Example 123

d) Adjective word order. The default word order of the adjective preceding the noun is not always kept in machine translation. This does not affect the meaning of the sentence but its fluency, as shown in Example 124. In this case, the role played by the N-gram Module and the LM Module is crucial, since the dependency parser can sometimes handle word order differences and analyse correctly those chunks even if the adjective follows the noun.

Example 124

HYP: *He said that in Spain "suspected of some **detainees Moroccans** clearly they participated directly or indirectly in preparation..."*

REF: *Bouzoubaa said that in Spain "some **Moroccan detainees** are clearly susptected of having directly or indirectly participated in the preparations..."*

After carrying out this qualitative analysis, it coincides with correlations with human judgements since both of them recommend the use of features related to morphology, morphosyntax, and syntax to evaluate the fluency of a segment.

Last but not least, it is worth mentioning the use of the LM Module. The LM model works as a complement of the reference translations, since those grammatical chunks not covered by the reference segments can be covered by the LM. This is the case of Example 125 where the use of the LM moves the score of the metric from 1.4 (using dependency and N-gram Modules) up to 2.5, coinciding with the human judgement for this segment.

Example 125

HYP: *He said the official, who asked to remain anonymous, "we support if the meeting is aimed at helping the Palestinian Authority at the level of economic and encourage them to undertake reforms".*

REF: *The official, who wished to remain anonymous, said "we support this meeting if the aim is to help the Palestinian Authority economically and to encourage it to make reforms".*

The Dependency Module accounts for the chunks:

- *to remain anonymous*
- *helping the Palestinian Authority*

The N-gram Module matches the chunks:

- *the official, who*
- *to remain anonymous, "we support*
- *helping the Palestinian Authority*
- *economic and*
- *encourage them to undertake reforms".*

In addition, by employing an LM we can account for the grammaticality of other chunks, such as *if the meeting is aimed at*, which were not covered by any of the previous modules because it does not occur in the reference sentence. Thus, using an

LM in combination with other modules aimed at checking the grammaticality of a segment turns into a positive contribution.

Next, the main findings yielded by these experiments are reported and discussed.

### 7.9 Findings on Fluency

This chapter has described the experiments carried out to check the suitability of both linguistic features and VERTa's modules to evaluate the fluency of a segment. The results of these experiments have also been reported and discussed and findings on a per-module basis have been yielded:

**Lexical Module**. Adapting the Lexical Module to the assessment of fluency does not imply important changes. However, it is noticeable the difference obtained in terms of correlation with human judgements. The Lexical Module itself achieved a high correlation with human judgements when experiments on adequacy were performed (see section 6.2). The reason for this high correlation is that the Lexical Module is mainly based on lexical semantics; thus focusing on meaning rather than grammaticality. As a consequence, correlation with human judgement drops dramatically when evaluating fluency because this module uses lexical items in isolation and no grammaticality features are covered here.

Regarding the linguistic information in this module, the same type of linguistic features used when assessing adequacy can be used when assessing fluency. It must be highlighted that although some of the features did not correlate well with human judgements (e.g. synonyms and partial lemma), they have been kept as part of the linguistic information used in this module. This decision was taken on the basis that the decrease in the correlation was not caused by those relations themselves, but other factors that could be solved by using other modules in VERTa (e.g., morphosyntax, n-grams and Dependency Modules). Besides, using such semantic relations favoured a wider coverage of lexical items, which otherwise would be disregarded. This corroborates our hypothesis that a linguistic analysis could clarify which linguistic features should be used to evaluate MT output. It is also noticeable the performance of hypernyms and hyponyms, that shows better results in these experiments than in those carried out for adequacy.

**Morphological Module**. Experiments conducted have shown that the combination of lexical information and PoS tags improves the correlation with human judgements on fluency. In addition, the use of PoS tags does not only help to identify problems in terms of agreement and verb tense, but also as regards missing information.

**Dependency Module**. As for the Dependency Module, the Exact match has shown to be the most appropriate for this type of evaluation, although correlations with human judgements indicated the opposite. The qualitative study performed has revealed that although the Exact match restricts the performance of this module, using more flexible types of matches (i.e. No_label, No_mod and No_head matches) only had positive effects in covering lexical semantic relations. In addition, more flexible matches became a problem because they allowed for matching constructions conveying similar meaning but which might not be completely grammatical, thus leading to a drawback in terms of fluency. The lack of more flexible matches was balanced by the use of language-dependent rules that aim at comparing two grammatically-correct syntactic constructions expressing the same meaning.

It must be noticed that using only the Exact match results in a rather mean performance of the metric, in the sense that it correlates well with low human judgements but, when human judgements assign higher scores VERTa has problems to reach those scores, especially due to the fact that the No_mod and No_head match, allowing for a wider coverage of lexical semantic relations are disregarded, as mentioned in the previous paragraph. Thus, looking for other ways to account for these lexical semantic relations in the Lexical Module would be advisable.

In addition, identifying instances of No_head match and No_mod match might be interesting in the error analysis field, since they help in identifying untranslated words or incorrect translations. Actually, the No_head match tends to account for untranslated or incorrect translations of verbs and nouns, whereas the No_mod match tends to account for untranslated or wrongly translated elements such as prepositions or determiners. This confirms our hypothesis that organising information at different levels and aimed at different tasks might help to detect MT errors.

Finally, organising dependency labels into different groups (top nodes, middle nodes and ultimate nodes) and assigning different weights to each group (the higher the node, the higher the weight) has also proved effective since somehow these weights reflect the importance of those dependency relations in the grammaticality of the sentence.

The different application of the Dependency Module depending on the type of evaluation performed, either adequacy (see section 6.4) or fluency (see section 7.4) confirms our hypothesis that depending on how syntactic information is used it can account for different types of evaluation.

**N-gram Module**. Regarding the N-gram Module, the best correlations are obtained when n-grams are calculated over PoS, since several features related to grammaticality are combined: morphology, morphosyntax and word order. This corroborates our hypothesis that that morphology (i.e. lemma and PoS), morphosyntax and word order, together with dependency relations, seems to be the most convenient to evaluate fluency**.** Moreover, since English word order is rather fixed, longer n-grams have proved more effective than shorter ones, thus accounting for longer sequences of correct word order.

**Semantic Module**. The Semantic Module has also been tested, however, as expected its influence on the fluency of a segment is rather low; actually, most of the semantically-motivated metrics obtained a negative correlation, except for NEL which seems to account for NEs that have been mistranslated or whose translation is not the one reported in Wikipedia.

**LM Module**. The language module did not show an outstanding performance when used in isolation, but it turned very useful when the modules were combined. The fact that the LM does not rely on reference translations opens a new path to cover language instances that are not reflected in the reference translations, thus allowing for a wider coverage. Apart from testing its contribution, one more experiment was conducted to test the suitability of three different LMs. From results obtained, it is clear that the LM used has to be domain-related in order to have a stronger influence and that a PoS-based LM might also be more appropriate to evaluate the fluency of a segment.

Finally, after analysing all modules separately, it is clear that those that correlate best with human judgements on fluency are the Dependency, and N-gram Modules. In addition, automatic tuning of weights was calculated and the ideal combination seems to be: Dependency Module (0.37), LM Module (0.30), N-gram Module (0.29) and Morphological Module (0.04). The Morphological Module plays a minor role mainly because a) English does not show a rich inflectional morphology, and b) PoS information is already used in the N-gram Module, covering morphosyntactic issues such as subject-verb agreement. This confirms our hypothesis that the information that should be combined to check the grammaticality of a sentence, thus its fluency, should combine: morphosyntactic information (e.g. lemma, PoS), word order and dependency relations.

Chapter 6 and Chapter 7 have covered the experiments to test the suitability and combination of linguistic information to evaluate both adequacy and fluency and VERTa has been the tool used to perform those experiments as well as the quantitative and qualitative analyses. Up to now, VERTa has proved effective as a tool, but now it is our interest to evaluate VERTa as an MT metric. With this aim in mind, next chapter presents a meta-evaluation of VERTa comparing it with other well-known and widely-used MT metrics.

# Chapter 8. Meta-Evaluation of VERTa

VERTa was first developed as a mere tool to explore and analyse the most appropriate and effective linguistic information required to perform MT evaluation and how this information should be combined; however, VERTa has another function: it can also be used as an MT metric to evaluate MT output. With the aim of confirming its validity as an MT metric, a meta-evaluation of VERTa has been performed. This meta-evaluation has covered three main areas following both a quantitative and a qualitative approach: adequacy, fluency and ranking (see section 2.1.2 for further details on each type of evaluation). This chapter covers the meta-evaluation of VERTa and is organised as follows: section 8.1, deals with the meta-evaluation of VERTa to test adequacy, compares its performance with that of other well-known metrics and provides both a quantitative and a qualitative analysis; section 8.2 covers the meta-evaluation of VERTa to evaluate fluency, compares its performance to that of other well-known metrics and presents both a qualitative and a quantitative analysis; finally, section 8.3 presents VERTa's participation at the ACL-WMT14, where VERTa competed with other metrics in a shared task aimed at ranking segments and systems translating from other languages to English.

## 8.1 Meta-Evaluation of VERTa to Test Adequacy

VERTa was mainly developed to explore the most relevant linguistic features to evaluate adequacy and fluency in MT output. To this aim several experiments were performed (see chapters 6 and 7). After these experiments, we were particularly interested in comparing VERTa to other well-known metrics. In order to carry a meta-evaluation on adequacy, the unseen part of the news-related corpus (Arabic-English) described in section 3.2.1.1 was used. This corpus contains 149 segments translated by 8 different systems, 4 reference translations and adjusted human judgements for adequacy. In order to check VERTa's performance, the same corpus has also been evaluated by several other metrics[44] contained in the Asiya framework[45] (Giménez and Márquez 2010a; González and Giménez 2014):

[44] For further details on the metrics presented here, please refer to section 2.2.2. Here we just provide a brief account of their functioning.

- BLEU: accumulated BLEU score up to 4-grams.

- METEOR-ex, METEOR-st, METEOR-sy and METEOR-pa: from using only exact matching (METEOR-ex), adding stem matching (METEOR-st), plus synonymy matching (METEOR-sy), plus paraphrase matching (METEOR-pa).

- SP-Op(*) and SP-Oc(*): metrics using shallow parsing. SP-Op(*) calculates the average lexical overlap over PoS tags. SP-Oc(*) calculates the average lexical overlap over all chunk types.

- DPm-Ol(*), DPm-Oc(*) and DPm-Or(*). These measures capture similarities between dependency trees in the hypothesis and reference segments and use the MALT v1.7 parser to analyse the segments. DPm-Ol(*) calculates overlapping between words hanging at all levels, DPm-Oc(*) calculates overlapping between grammatical categories, and finally, DPm-Or(*) calculates overlapping between grammatical relations.

- CP-Op(*) and CP-Oc(*)[46]. These measures compare similarities between constituent parse trees in the hypothesis and reference segments. The Charniak and Johnson (2005)'s Max-Ent reranking parser is used to obtain the constituent trees. CP-Op(*) calculates lexical overlap over PoS and CP-Oc(*) calculates lexical overlap according to the phrase constituent.

- SR-Or, SR-Or(*) and SR-Mr(*). These metrics compare Semantic Roles similarities between the hypothesis and reference segments. SR-Or deals with Semantic Roles overlap regardless of their lexical realization. SR-Or(*) computes the average lexical overlap over all Semantic Roles types. SR-Mr(*) calculates the average lexical matching over all Semantic Roles types.

- NE-Me(*) and NE-Oe(*). This set of metrics compares the hypothesis and reference segments according to their NEs. The NE-Me(*) calculates the

[45] http://asiya.lsi.upc.edu/

[46] Although both SP and CP metrics use the Penn Treebank PoS tagset, SP metrics use a different tool to automatically annotate sentences (SVM tool (Giménez and Márquez 2004) and BIOS (Surdeanu et al. 2005) ), thus its different performance.

average lexical matching over all NEs whereas the NE-Oe(*) calculates the average lexical overlap over NEs.

- Combination of metrics 1: The ULC (Unified Linear Combination) combination of metrics that are representative of each linguistic level in Asiya (Giménez and Márquez 2008b). This set of metrics includes: BLEU, NIST, -TER, -TERp-A, ROUGE-W, METEOR-ex, METEOR-pa, METEOR-st, METEOR-sy, DP-HWCM_c-4, DP-HWCM_r-4, DP-Or(*), CP-STM-4, SR-Or(*), SR-Mr(*), SR-Or, DR-Or(*), DR-Orp(*). They are combined by means of the normalized arithmetic mean of all metrics' scores.

- Combination of metrics 2: The ULC combination of metrics that according to Giménez and Márquez (2010b) show the best performance in several data sets to evaluate quality. This combination of metrics is: ROUGE-W, METEOR-sy, DP-HWCM_c-4, DP-HWCM_r-4[47], DP-Or(*), CP-STM-4, SR-Or(*), SR-Mr(*), SR-Or, DR-Or(*), DR-Orp(*).

Modules in VERTa have been set and combined according to the results obtained from the experiments on adequacy (see Chapter 6). Thus the modules used and weights assigned are the following: Lexical Module (0.47), Dependency Module (0.43), N-gram Module (0.05) and Semantic Module (0.05).

Correlations with human judgements obtained by these metrics have been compared to the correlation obtained by VERTa and both a quantitative and a qualitative analysis of the results has been conducted.

### 8.1.1 Results and Discussion

This section presents the correlations obtained by the above mentioned metrics and compares them to that obtained by VERTa. A thorough analysis of these results has also been conducted and both quantitative and qualitative analyses have been performed.

[47] In the original combination of metrics, there were two metrics that are not available in the Asiya framework nowadays, DP-HWCM_c and DP-HWCM_r, and which have been substituted by the variants DP-HWCM_c-4 and DP-HWCM_r-4.

Table 75 shows the Pearson correlation obtained by the metrics described above and by VERTa.

| Metric | Pearson Correlation |
|---|---|
| VERTa | **0.7289** |
| BLEU | 0.5771 |
| METEOR-ex | 0.5687 |
| METEOR-st | 0.5715 |
| METEOR-sy | 0.5690 |
| METEOR-pa | 0.5521 |
| SP-Oc(*) | **0.6292** |
| SP-Op(*) | 0.5707 |
| DPm-Ol(*) | 0.5430 |
| DPm-Oc(*) | 0.2685 |
| DPm-Or(*) | 0.5748 |
| CP-Op(*) | **0.6166** |
| CP-Oc(*) | **0.6292** |
| SR-Or(*) | 0.3925 |
| SR-Mr(*) | 0.3079 |
| SR-Or | 0.1825 |
| NE-Me(*) | 0.3040 |
| NE-Oe(*) | 0.3325 |
| Metric Combination 1 | **0.6506** |
| Metric Combination 2 | 0.5788 |

**Table 75** Pearson correlation for adequacy. Comparing VERTa metric and a selection of well-known metrics

#### 8.1.1.1 Quantitative Analysis

According to the results obtained, VERTa stands out from the rest of the metrics obtaining a correlation of 0.7289, whereas the closest metrics get 0.6506 (Combination 1), 0.6292 (SP-Oc/CP-Oc metrics) and 0.6166 (CP-Op). The key factor for VERTa's excellent performance is the combination of linguistic information at different levels that enriches the metric and allows for a more flexible use.

A closer analysis of the results shows that those metrics working at lexical level (BLEU and METEOR family) obtain similar results. It is interesting to notice that in the METEOR family, the more linguistic information used, the worse the correlation obtained. The only type of information which improves its correlation is the use of stemming, however the use of synonymy has the opposite effect. This is quite surprising since adequacy is being evaluated, thus the use of synonymy relations seemed to be appropriate. Actually, the use of synonyms in VERTa has proven to be effective in order to increase its correlation with human judgements (see section 6.2.1). Finally, the version of METEOR using paraphrasing obtains the lowest score among the metrics family, which might indicate that paraphrasing produces noise which impoverishes the performance of the metric.

Regarding those metrics using syntactic information, their performance seems to contradict the common belief that this type of metrics is the most effective one to evaluate the fluency of a segment (and thus not recommending their use for the evaluation of adequacy), since some of them (SP-Oc, CP-Oc and CP-Op) obtain a good correlation with human judgements of adequacy. It is noticeable that those that work at chunk and phrase constituency level are the ones that obtain the best results. Hence, this seems to indicate that word order is also important when evaluating adequacy, confirming the modules combination in VERTa, where the N-gram Module also proved effective. On the other hand, those metrics working with dependency trees do not obtain good results. Actually, within the DPm family, the metric obtaining the lowest correlation is the metric calculated over grammatical categories (DPm-Oc(*)), which gets 0.2685. This was quite expected, since grammatical categories are far from dealing with meaning. On the other hand, both DPm-Ol (0.5430) and DPm-Or (0.5748) show a better performance than DPm-Oc (0.2685) because they compare lexical items, the former, and dependency relations, the latter. The low performance of both the DPm-metrics family in comparison with VERTa's Dependency Module might be due to the fact that both metrics families are much more rigid than VERTa. The key factors for VERTa's better performance are that a) in VERTa's Dependency Module, information regarding lexical semantics has also been taken into account; b) VERTa's Dependency Module considers different types of matches and rules which lead to a more flexible coverage of dependency relations and allows for similarity between different syntactic

structures conveying the same meaning, even if they are not totally grammatical; c) in VERTa, the least informative dependency relations are assigned very low weights. Another factor that might also be taken into account when comparing VERTa's Dependency Module and the DPm family is the selection of the dependency parser used to perform the analysis (see Comelles et al. (2010)'s paper on evaluating constituency and dependency parsers); VERTa uses the Standford parser, whereas the DPm family makes use of the MALT parser.

Finally, regarding those metrics more related to semantics – SR-based metrics and NEs-based metrics – they did not obtain a good correlation. Actually, a better performance was expected, especially from those using Semantic Roles information. According to Lo and Wu (2010), this type of information is especially useful when evaluating adequacy, however, results obtained by the SR-metrics contradict their statement. It must be noticed that those metrics that compare Semantic Roles taking into account lexical items - SR-Or(*) (0.3925) and SR-Mr(*) (0.3079) - work better than that which disregards their lexical realization (SR-Or), which gets 0.1825. The reason behind may be that the latter does not distinguish the structural relations established within semantic frames. Another possibility for the low correlation of such metrics might be the performance of the tool used for Semantic Role labelling. Last but not least, the NEs-based metrics obtained a low correlation (0.3040 for the NE-Me(*) and 0.3325 for the NE-Oe(*)), similar to those obtained by NE-based components in VERTa. These results were expected since, as explained in section 6.6.1, NEs are just a partial aspect of the segment and human judgements used for correlation assess the hypothesis segment as a whole. However, it must be noticed that even though these metrics do not correlate well in isolation, they slightly contribute when combined with other modules.

In addition, since VERTa combines linguistic features at different levels, two combinations of some of the metrics available in Asiya have also been used in order to compare VERTa to other combinations of metrics. Results for combination 1[48] confirm our hypothesis that the combination of several metrics working at different levels correlate better with human judgements than single metrics working at a specific level. On the other hand, according to the correlations obtained, VERTa outperforms

[48] The one obtaining the best results between the two.

significantly Combination 1, which gets 0.6506. This is mainly due to the fact that VERTa's individual modules are more flexible and use more linguistic information than those in this combination. In addition, it must also be highlighted that metrics in combination 1 are combined using the normalized arithmetic mean of all metrics scores, whereas VERTa selects and weighs each module depending on the type of evaluation. Finally, Combination 1 uses a wide range of metrics so it is difficult to check the influence of each metric and whether any of them represents a drawback to this type of evaluation. Thus, it seems that the combination of such a large amount of metrics is not that effective and a selection of metrics covering those key linguistic features related to the meaning of a sentence (e.g. those in VERTa) have proved more effective to evaluate adequacy.

#### 8.1.1.2 Qualitative Analysis

Once correlations were compared, a qualitative analysis of those scores was performed. This analysis has confirmed that VERTa obtains better results than those metrics referred to in Table 75 because it is more flexible than other metrics.

This flexibility is first seen in the use of synonyms, lemma and partial lemma features, as shown in the example below. This is a key feature to distinguish VERTa's performance from most of the metrics reported in Table 75, except for METEOR and ULC-Combination1 that also use information related to synonymy and stemming.

Example 127

HYP: *...when he said "superiority of Western culture"...*

REF: *...when he stated that "western civilization is superior"...*

In Example 127, *said* and *stated* are matched thanks to the use of synonymy relations and *superiority* and *superior* thanks to partial lemma. In addition, the use of flexible dependency matches allows for matching *superiority of Western culture* and *western civilization is superior* which convey the same meaning, as illustrated by the No_label match in Table 76. Likewise, this flexibility of matches also allows for comparing the direct object in the hypothesis sentence with the clause complement in the reference segment by employing the No_label match.

| Hypothesis | Reference | Match |
|---|---|---|
| advmod(said,when) | advmod(stated,when) | Exact match |
| nsubj(said,he) | nsubj(stated,he) | Exact match |
| **dobj(said,superiority)** | **ccomp(said,superior)** | **No_label match** |
| amod(culture,Western) | amod(civilization,western) | Exact match |
| **prep_of(superiority,culture)** | **nsubj(superior,civilization)** | **No_label match** |

**Table 76** Dependency matches for Example 127

Another example of the benefits that the flexibility of the Dependency Module provides is Example 128.

Example 128

HYP: *He said that "unknown gunmen killed Lieutenant Uday Khyoun at 00,19 local time".*

REF: *He said that "unidentified gunmen killed Lieutenant Udai Khayoun at 1900 local time".*

| Hypothesis | Reference | Match |
|---|---|---|
| nsubj(said,He) | nsubj(said,He) | Exact match |
| root(TOP,said) | root(TOP,said) | Exact match |
| complm(killed,that) | complm(killed,that) | Exact match |
| amod(gunmen,unknown) | amod(gunmen,unidentified) | Exact match |
| nsubj(killed,gunmen) | nsubj(killed,gunmen) | Exact match |
| ccomp(said,killed) | ccomp(said,killed) | Exact match |
| nn(Khyoun,Lieutenant) | nn(Khayoun,Lieutenant) | No_head match |
| nn(Khyoun,Uday) | NO MATCH | NO MATCH |
| dobj(killed,Khyoun) | dobj(killed,Khayoun) | No_mod match |

**Table 77** Dependency matches from Example 128

Most of the matches in Table 77 are Exact matches, except for a No_mod match and No_head match which allow for comparing proper nouns which have not been translated exactly the same way. METEOR does not allow for these matches, as well as

other metrics working at syntax level which do not use synonyms, lemma and partial lemma information.

After focusing on the meta-evaluation of VERTa on adequacy, next section provides information regarding the meta-evaluation on fluency.

### 8.2. Meta-Evaluation of VERTa to Test Fluency

Once VERTa has been evaluated on adequacy, it is the turn for fluency. To this aim, the unseen part of the news-related corpus (Arabic-English) described in section 3.2.1.2 has been used. This corpus contains 149 segments translated by 6 different systems, 4 reference translations and human judgements on fluency per segment provided by 2 different judges. In order to obtain a single judgement per segment, the average was calculated. VERTa's performance was compared to well-known metrics such as BLEU, the METEOR family and some of the linguistically-based metrics available in the Asiya framework. Most of these metrics have been already described in section 8.1, however others have been added because they were more fluency-oriented[49]. These are:

- DP-HWCM_c-4 and DP-HWCM_r-4 metrics, variants of Liu and Gildea (2005)'s HWCM metric which consider different head-word chain types. DP-HWCM_c-4 considers syntactic categories whereas DP-HWCM_r-4 considers syntactic relations and both of them calculate the average accumulated proportion of category/relation chains up to length 4;

- Confidence Estimation (CE) measures (Specia et al. 2010) also available in the Asiya framework that seem suitable to check the fluency of a segment. CE measures do not need reference translations, they can be target-based (just focusing on target segments) or source/target-based (using both source and target sentences). From those CE measures available in Asiya we selected three target-based measures since their hypothesis is that the likelier the sentence (according to a language model), the more fluent. Hence they are suitable in order to check the fluency of a segment. These three measures are: CE-ippl, CE-ippl-c and CE-ippl-p. CE-ippl calculates the inverse perplexity of the target segment according to a pre-defined language model. CE-ippl-c metric combines

[49] As for adequacy, for a full description of these metrics, please refer to Section 2.2.2.

the use of a language model with phrase chunks tags. Finally, CE-ippl-p metric uses a language model calculated over sequences of PoS tags. For further details please refer to Asiya technical manual (González and Giménez 2014).

### 8.2.1. Results and Discussion

This section presents a comparison between VERTa and the metrics described above when used to evaluate the fluency of a segment and offers both a quantitative and a qualitative evaluation of the results obtained.

Table 78 reports results obtained by VERTa and the selected set of well-known metrics, when comparing their scores to human judgements on fluency by means of Pearson correlation coefficient.

#### 8.2.1.1 Quantitative Analysis

VERTA's correlation with human judgements on fluency is worse than the correlation obtained with adequacy judgements. This was not unexpected since similar results were obtained when the experiments were performed (section 7.8). Nonetheless, it must be noticed that VERTa clearly outperforms the metrics it is compared against.

Although BLEU is one of the widest used metrics to evaluate MT quality and has also been claimed to correlate well with human judgements on fluency, it has not proved effective to evaluate fluency with our data. This was somehow anticipated given the strict word order considered by BLEU and its matches. As for metrics in the METEOR family, they got similar results, although METEOR-sy, which covers exact matches, stemming and synonymy relations, obtains the best correlation (0.3275). It might be due to the fact that METEOR-sy allows for matching a wider range of lexical items which results into a lower penalisation for unmatched items.

| Metric | Pearson Correlation |
|---|---|
| VERTa | **0.4553** |
| BLEU | 0.2933 |
| METEOR-ex | 0.3086 |
| METEOR-st | 0.3079 |
| METEOR-sy | 0.3275 |
| METEOR-pa | 0.3182 |
| SP-Oc(*) | 0.3113 |
| SP-Op(*) | 0.2849 |
| DPm-Ol(*) | 0.2374 |
| DPm-Oc(*) | 0.2471 |
| DPm-Or(*) | 0.3578 |
| DP-HWCM_c-4 | 0.2505 |
| DP-HWCM_r-4 | 0.2483 |
| CP-Op(*) | 0.3158 |
| CP-Oc(*) | **0.3686** |
| SR-Or(*) | 0.3040 |
| SR-Mr(*) | 0.2379 |
| NE-Me(*) | 0.0806 |
| NE-Oe(*) | 0.0728 |
| CE-ippl | 0.1932 |
| CE-ippl-c | 0.1461 |
| CE-ippl-p | 0.2079 |

**Table 78** Pearson correlation for fluency. Comparing VERTa and a selection of well-known metrics

From those metrics using shallow parsing, SP-Oc(*) and SP-Op(*), the former, which accounts for all successfully translated phrases, achieves good results (0.3113). This is due to the fact that the metric checks that all words inside a specific phrase have been translated correctly and, indirectly, it accounts for correct word order inside the phrase. As for metrics using dependency trees information, DPm-Or(*) shows a good correlation (0.3578) in line with the Dependency Module in VERTa. Such good results in comparison to DPm-Ol (0.2374) and DPm-Oc (0.2471) are due to the fact that DPm-

Or computes overlap between words ruled by non-terminal nodes, thus covering both syntactic relations at phrase level occupying higher positions in the syntactic tree. On the other hand, although Liu and Gildea (2005) claimed that their HWCM metric achieved good results as regards fluency, that is not the case with the two variants tested DP-HWCM_c-4 (0.2505) and DP-HWCM_r-4 (0.2483).

As for metrics working at constituent level, the CP-Oc(*) metric obtains the best correlation from all metrics used (0.3686), except for VERTa. Without doubt, the use of syntactic information on constituents, namely lexical overlap according to the phrase constituent, proves effective to evaluate the fluency of a segment; thus, highlighting again the importance of word order, not only in phrase chunks but most importantly inside phrase constituents to check the grammaticality of a sentence.

On the other hand, as expected, semantically-based metrics do not achieve good correlations with human judgements on fluency. This is especially remarkable in NE metrics which show a very poor performance (0.0806 for Ne-Me(*) and 0.0728 for NE-Oe(*)), in line with VERTa's NE components.

A new set of metrics has been used to evaluate fluency, CE metrics, namely CE-ippl, CE-ippl-c and CE-ippl-p. Unfortunately, none of them obtain a good correlation, being CE-ippl-p the best one (0.2079). This might be due to the LM used, based on the Europarl corpus, a different genre from the newswire corpus used to conduct this meta-evaluation and to the fact that they were used isolated. This reinforces the idea that LMs are more valuable in domain-restricted contexts. On the other hand, from this set of metrics, CE-ippl-p obtained the best results. This metric uses an LM calculated over sequences of PoS tags, which reinforces the idea that PoS tags and word order are appropriate to evaluate fluency and that LM-based measures contribute in the evaluation when combined with other information, as shown in results reported in section 7.8.

To conclude, according to the results obtained, a collaborative approach such as that proposed in VERTa, which combines information on dependency relations, PoS tags and word order, is the most appropriate to evaluate the grammaticality of a sentence. The combination of different linguistic features, once again, outperforms single metrics.

#### 8.2.1.2 Qualitative Analysis

After analysing the results obtained quantitatively, a more qualitative analysis has been conducted so as to compare VERTa to CP-Oc(*) and DPm-Or(*), the two metrics that obtained the best results after VERTa. The most relevant points are detailed below.

Firstly, the use of synonymy relations in VERTa allows for a more flexible match. That is the case of Example 129, where the hypothesis string is completely grammatical.

Example 129

HYP: *...a new **stage** after the death of Palestinian Authority Yasser Arafat ...*

REF: *...a new **phase** "after the death of Palestinian Authority leader Yasser Arafat...*

Since the Dependency and the Morphologocial Modules in VERTa make use of synonyms, it is able to match *stage* and *phase*, whereas neither CP-Oc(*) nor DPm-Or(*) can establish such a similarity because they do not use synonymy relations. In this sense, CP-Oc(*) penalises a chunk which in fact is absolutely fluent.

Secondly, the use of language-dependent rules in VERTa's Dependency Module helps to compare two grammatically-correct constructions that convey the same meaning, even if they show a different syntactic structure. Example 130 shows the equivalent between chunks *leaders of Hong Kong* and *Hong Kong's leaders*, two equivalence and grammatical expressions that are realised by two different grammatical structures (see Table 79). Since both of them are grammatically correct, they must be considered similar and as a positive match in terms of fluency. Both CP-Oc(*) and DPm-Or(*) metrics fail in finding such a similarity since they are realised by two different grammatical relations and phrase types.

Example 130

HYP: *...criticism to **the leaders of Hong Kong**...*

REF: *...criticism at **Hong Kong's leaders**...*

| Hypothesis | Reference | Match |
|---|---|---|
| nn(Kong,Hong) | nn(Kong,Hong) | Exact match |
| prep_of(leaders,Kong) | poss(leaders,Kong) | [prep_of]-[poss] match |

**Table 79** Dependency match similarity for Example 130

Finally, the use of N-grams over PoS matches, instead of lexical matches, has also proved effective since it has helped to restrict the wide coverage of the Lexical Module, as illustrated in Example 131.

Example 131

HYP: *...you will be more* ***unity-NN*** *more* ***cooperation-NN****...*

REF: *...you must be more* ***united-VBN*** *and more* ***cooperative-JJ****...*

If N-gram matches were based on lexical matches, both *unity* - *united* and *cooperation* - *cooperative* matches would be considered positive matches by using the partial lemma match. However, since the N-gram Module works over PoS, this match has been disregarded.

Sections 8.1 and 8.2 have dealt with the meta-evaluation of VERTa on the two tasks it was developed for: fluency and adequacy. Next section presents the evaluation of VERTa for a new task, MT quality using ranking of segments, in the context of the WMT14 Metrics Shared Task.

### 8.3 VERTa's Participation in WMT14: MT Quality Using Ranking

The Workshop on Statistical Machine Translation is one of the most prestigious venues for research in computational linguistics in general, and Machine Translation in particular. In this workshop several Shared Tasks are held, specifically WMT14 has held a Translation Task, a Metrics task, a Quality Estimation Task and a Medical Translation Task.

The Shared Metrics Task examines MT evaluation metrics with the aim of achieving the strongest correlation with human judgements of translation quality. It must be highlighted that human judgements are not based on either adequacy or fluency, but on translation quality as a whole. In addition, human judgements are based on sentence

ranking, in other words, for each source sentence human judges are provided with the outputs of five systems to which they have to assign ranks. Ties are allowed. All this poses a new challenge to VERTa since it has not been developed to deal with translation quality as a whole, but with adequacy and fluency separately. Moreover, VERTa is not a ranking metric but it provides scores for each segment evaluated.

### 8.3.1 Preliminary Experiments

Although the time and computational resources that we had available were limited and rather scarce, some preliminary experiments were conducted on previous WMT editions' data, specifically on WMT12, WMT13, all languages into English (en) (see section 3.2.1.3). Languages "all" include French (fr), German (de), Spanish (es) and Czech (cz) for WMT12; and French, German, Spanish, Czech and Russian (ru) for WMT13. In both campaigns only 1 reference was used. Data sets distributed are reported in Tables 80 and 81, respectively.

| WMT12 Data | cs-en | de-en | fr-en | es-en | Total |
|---|---|---|---|---|---|
| **#systems** | 6 | 16 | 15 | 12 | 49 |
| **#segments per system** | 3,003 | 3,003 | 3,003 | 3,003 | 12,012 |
| **#segments** | 18,018 | 48,048 | 45,045 | 36,036 | **147,147** |

**Table 80** WMT12 data

| WMT13 Data | cs-en | de-en | fr-en | ru-en | es-en | Total |
|---|---|---|---|---|---|---|
| **#systems** | 12 | 23 | 19 | 23 | 13 | 90 |
| **#segments per system** | 3,000 | 3,000 | 3,000 | 3,000 | 3,000 | 15,000 |
| **#segments** | 36,000 | 69,000 | 57,000 | 69,000 | 39,000 | **270,000** |

**Table 81** WMT13 data

Both segment and system level evaluations were performed. Evaluation sets provided by WMT organizers were used to calculate both segment and system level correlations, which included: Kendall's tau correlation coefficient (Kendall 1938/1955) at segment level and Spearman's correlation coefficient (Spearman 1904) at system level.

Since VERTa has been mainly designed to assess either adequacy or fluency separately, our goal for WMT14 was to find a rather effective combination of modules in order to

evaluate translation quality in general. Firstly, we decided to explore the influence of each module separately. To this aim, all modules were used, except for the Semantic and LM Modules, which were not available at that time. Secondly, all modules were assigned the same weight and tested in combination (VERTa-EQ). Intra-module features were set as if adequacy was tested, to allow more flexible types of matches. Thus, each module was set as follows:

- Lexical Module. As described in Chapter 6, section 6.2.4, except for the use of hypernyms/hyponyms matches that were disregarded.
- Morphological Module. As described in Chapter 6, section 6.3.3, except for the lemma-PoS match and the hypernyms/hyponyms-PoS match.
- Dependency Module. As described in Chapter 6, section 6.4.5.
- N-gram Module. As described in Chapter 6, section 6.5.3, using a 2-gram length.

Experiments aimed at evaluating the influence of each module (see Table 82 and Table 83) show that the Dependency Module, in the case of WMT12 data, and the Lexical Module, in the case of WMT13 data, are the most effective ones. However, the influence of the N-gram Module and the Morphological Module varies depending on the source language. The fact that the Dependency Module correlates better with human judgements than others might be due to its flexibility to capture different syntactic constructions that convey the same meaning. In addition, the good performance of the Lexical Module is due to the use of lexical semantic relations. On the other hand, in general the Morphological Module shows a better performance than the N-gram one on the WMT12 dataset, whereas this is not always true for the WMT13 dataset. Thus, it was difficult to decide which of the modules was the most useful. A thorough analysis based on the data rather than on the correlations obtained would have been advisable in order to obtain more information regarding intra-module features. However, due to the lack of time and computational resources this analysis could not be conducted.

| Module | fr-en | de-en | es-en | cs-en |
|---|---|---|---|---|
| **Lexical M** | .16 | .20 | .18 | .14 |
| **Morphological M.** | .17 | .19 | .18 | .12 |
| **Dependency M.** | **.18** | **.24** | **.20** | **.17** |
| **N-gram M** | .16 | .17 | .15 | .08 |

**Table 82** Segment-level Kendall's tau correlation per module with WMT12 data

| Module | fr-en | de-en | es-en | cs-en | ru-en |
|---|---|---|---|---|---|
| **Lexical M.** | **.239** | **.254** | **.294** | **.227** | **.220** |
| **Morphological M.** | .236 | .243 | .295 | .214 | .191 |
| **Dependency M.** | .232 | .247 | .275 | .220 | .199 |
| **N-gram M.** | .237 | .245 | .283 | .213 | .189 |

**Table 83** Segment-level Kendall's tau correlation per module with WMT13 data

Moreover, a second version of VERTa was employed (VERTa-W). This new version used the module combination aimed at evaluating adequacy, which is mainly based on the Dependency and Lexical Modules, but with a stronger influence of the N-gram Module in order to control word order (VERTa-W). Since English is not highly inflected and we wanted to provide a single version that could evaluate all languages into English, in VERTa-W the Morphological Module was disregarded in favour of the N-gram Module, which would help to account for word order. Weights for each module were manually assigned, based on results obtained in previous experiments conducted for adequacy and fluency (see Chapter 6, section 6.7 and Chapter 7, section 7.8), as follows:

- Lexical Module: 0.41
- Morphological Module: 0
- Dependency Module: 0.40
- N-gram Module: 0.19

Finally, the two versions of VERTa were compared: the unweighted combination (VERTa-EQ) and the weighted one (VERTa-W). These two versions were also

compared to some of the best performing metrics in WMT12 (see Table 84 and Table 85) and WMT13 (see Table 86 and Table 87): Spede07-pP (Wang and Manning 2012), METEOR (Denkowski and Lavie 2011), SEMPOS (Macháćek and Bojar 2011) and AMBER (Chen et al. 2012) in WMT12; SIMPBLEU-RECALL (Song et al. 2013), METEOR (Denkowski and Lavie 2011) and DEPREF-ALIGN (Wu et al. 2013) in WMT13 (Macháćek and Bojar 2013). As regards WMT12 data at segment level, the unweighted version achieves similar results to those obtained by the best performing metrics (see Table 84). On the other hand, VERTa-W's results are slightly worse, especially for fr-en (0.24) and es-en (0.25) pairs, which is probably due to the fact that the Morphological Module has been disregarded in this version. Regarding system level correlation, neither VERTa-EQ nor VERTa-W achieves a high correlation with human judgements (see Table 85).

| Metric | fr-en | de-en | es-en | cs-en | Average |
|---|---|---|---|---|---|
| **Spede07-pP** | **.26** | **.28** | **.26** | **.21** | **.25** |
| **METEOR** | .25 | .27 | .24 | .21 | .25 |
| **VERTa-EQ** | **.26** | **.28** | **.26** | .20 | .25 |
| **VERTa-W** | .24 | .28 | .25 | .20 | .25 |

**Table 84** Segment-level Kendall's tau correlation WMT12

| Metric | fr-en | de-en | es-en | cs-en | Average |
|---|---|---|---|---|---|
| **SEMPOR** | .80 | **.92** | .94 | **.94** | **.90** |
| **AMBER** | **.85** | .79 | **.97** | .83 | .86 |
| **VERTa-EQ** | .83 | .71 | .89 | .66 | .77 |
| **VERTa-W** | .79 | .73 | .91 | .66 | .77 |

**Table 85** System-level Spearman's rho correlation WMT12

As for segment level WMT13 results (see Table 86), although both VERTa-EQ and VERTa-W's performance is worse than that of the two best-performing metrics, both versions achieve a third and fourth position for all language pairs (0.261 in average for both), except for fr-en (0.252 and 0.253, respectively). As regards system level correlations (see Table 87), both versions of VERTa show the best performance for de-en (0.970 and 0.980) and ru-en (0.814 and 0.868) pairs, as well as for the average score (0.936 and 0.951).

| Metric | fr-en | de-en | es-en | cs-en | ru-en | Average |
|---|---|---|---|---|---|---|
| **SIMPBLEU-RECALL** | **.303** | **.318** | **.388** | **.260** | **.234** | **.301** |
| **METEOR** | .264 | .293 | .324 | .265 | .239 | .277 |
| **VERTa-EQ** | .252 | .280 | .318 | .239 | .215 | .261 |
| **VERTa-W** | .253 | .278 | .314 | .238 | .222 | .261 |
| **DEPREF-ALIGN** | .257 | .267 | .312 | .228 | .200 | .253 |

**Table 86** Segment-level Kendall's tau correlation WMT13

| Metric | fr-en | de-en | es-en | cs-en | ru-en | Average |
|---|---|---|---|---|---|---|
| **METEOR** | .984 | .961 | **.979** | **.964** | .789 | .935 |
| **DEPREF-ALIGN** | **.995** | .966 | .965 | **.964** | .768 | .931 |
| **VERTa-EQ** | .989 | **.970** | .972 | .936 | **.814** | **.936** |
| **VERTa-W** | .989 | **.980** | .972 | .945 | **.868** | **.951** |

**Table 87** System-level Spearman's rho correlation WMT13

Next, VERTa's participation at the WMT14 Metrics Shared Task is reported and discussed.

### 8.3.2 WMT14 Results and Discussion

In the WMT14 shared task, the data used was that provided by the organisation. These data contained systems' translations from French, German, Hindi (hi), Czech and Russian into English and 1 reference translation (see section 3.2.1.4 for a detailed description). The number of systems per language pair, number of segments per language pair, and the total number of segments are presented in Table 88.

| | cs-en | de-en | hi-en | fr-en | ru-en | Total |
|---|---|---|---|---|---|---|
| **#systems** | 5 | 13 | 9 | 8 | 13 | 48 |
| **#segments per syst.** | 3,003 | 3,003 | 3,003 | 3,003 | 3,003 | 15,015 |
| **#segments** | 15,015 | 39,039 | 27,027 | 24,024 | 39,039 | **144,144** |

**Table 88** Data provided in the WMT14 Shared Task, from all languages to English

This year Pearson correlation coefficient was used to calculate system-level correlations with human judgements, whereas Kendall's tau correlation coefficient was used to

calculate segment-level correlations. 22 metrics participated at system level and 18 at segment level.

Both versions of VERTa, VERTa-EQ and VERTa-W, were sent to the WMT14 task. As previously mentioned, since we did not have enough time and computational resources available, we could not conduct a detailed analysis to better adapt VERTa to the evaluation of MT quality in general and to the ranking task in particular. Thus, we were aware that we were not sending the best possible versions to this task. Surprisingly, though, at system level VERTa-W and VERTa-EQ occupied the 5$^{th}$ (0.906) and 6$^{th}$ (0.904) positions out of 22 and scored above average in all language pairs, which was a good result (see Table 89).

On average, VERTa was 0.04 points below the best metric DiscoTK-Party-Tuned (Joty et al. 2014), which scored 0.944. This metric uses information about discourse analysis in combination with 18 other metrics working at different levels (see section 2.2.2.4 for further details) and has been tuned using a learning-to-rank framework. The metric occupying the second position (0.927), Layered (Gautam and Bhattacharyya 2014), also uses a combination of metrics at different levels (i.e. lexical, syntax and semantics) and SVM rank to learn the appropriate parameters. The third position (0.918) in the ranking is occupied by the untuned version of DiscoTK-Party, and finally, UPC-Stout (González et al. 2014b) occupies the fourth position (0.913). UPC-Stout combines 32 different metrics, covering metrics that do not use linguistic information, metrics that use shallow parsing, constituent parsing, dependency parsing, Semantic Roles, NEs and source-based metrics. Both VERTa-W and VERTa-EQ combine less information than the metrics described above (DiscoTK-Party, Layered and UPC-Stout), it has not been developed for ranking of segments and in opposition to DiscoTK-Party-Tuned and Layered, no tuning of weights has been performed, their weight distribution relies only on linguistic grounds.

At segment level, VERTa-W and VERTA-EQ's performance decreased and on average both metrics occupied the 10$^{th}$ and 11$^{th}$ positions out of 18, respectively (see Table 90). However, both metrics scored above average in all language pairs (0.337 and 0.336, respectively).

| Metrics | Pearson Correlation Coefficient | | | | | |
|---|---|---|---|---|---|---|
| | fr-en | de-en | hi-en | cs-en | ru-en | Average |
| DISCOTK-PARTY-TUNED | .977 | .942 | .956 | .975 | .870 | .944 |
| LAYERED | .973 | .893 | .976 | .941 | .854 | .927 |
| DISCOTK-PARTY | .970 | .921 | .862 | .983 | .856 | .918 |
| UPC-STOUT | .968 | .915 | .898 | .948 | .837 | .913 |
| **VERTa-W** | **.959** | **.867** | **.920** | **.934** | **.848** | **.906** |
| **VERTa-EQ** | **.959** | **.854** | **.927** | **.938** | **.842** | **.904** |
| TBLEU | .952 | .832 | .954 | .957 | .803 | .900 |
| BLEU NRC | .953 | .823 | .959 | .946 | .787 | .894 |
| BLEU | .952 | .832 | .956 | .909 | .789 | .888 |
| UPC-IPA | .966 | .895 | .914 | .824 | .812 | .882 |
| CDER | .954 | .823 | .826 | .965 | .802 | .874 |
| APAC | .963 | .817 | .790 | .982 | .816 | .874 |
| REDSYS | .981 | .898 | .676 | .989 | .814 | .872 |
| REDSYSSENT | .980 | .910 | .644 | .993 | .807 | .867 |
| NIST | .955 | .811 | .784 | .983 | .800 | .867 |
| DISCOTK-LIGHT | .965 | .935 | .557 | .954 | .791 | .840 |
| METEOR | .975 | .927 | .457 | .980 | .805 | .829 |
| TER | .952 | .775 | .618 | .976 | .809 | .826 |
| WER | .952 | .762 | .610 | .974 | .809 | .821 |
| AMBER | .948 | .910 | .506 | .744 | .797 | .781 |
| PER | 0946 | .867 | .411 | .833 | .799 | .781 |
| ELEXR | .971 | .857 | .535 | .945 | -.404 | .581 |

**Table 89** System-level correlations of automatic evaluation metrics and the official WMT human scores when translating into English

| Metrics | Kendall's tau Correlation Coefficient | | | | | |
|---|---|---|---|---|---|---|
| | fr-en | de-en | hi-en | cs-en | ru-en | Average |
| DISCOTK-PARTY-TUNED | .433 | .380 | .434 | .328 | .355 | .386 |
| BEER | .417 | .337 | .438 | .284 | .333 | .362 |
| REDCOMBSENT | .406 | .338 | .417 | .284 | .336 | .356 |
| REDCOMBSYSSENT | .408 | .338 | .416 | .282 | .336 | .356 |
| METEOR | .406 | .334 | .420 | .282 | .329 | .354 |
| REDSYSSENT | .404 | .338 | .386 | .283 | .321 | .346 |
| REDSENT | .403 | .336 | .383 | .283 | .323 | .345 |
| UPC-IPA | .412 | .340 | .368 | .274 | .316 | .342 |
| UPC-STOUT | .403 | .345 | .352 | .275 | .317 | .338 |
| **VERTa-W** | **.399** | **.321** | **.386** | **.263** | **.315** | **.337** |
| **VERTa-EQ** | **.407** | **.315** | **.384** | **.263** | **.312** | **.336** |
| DISCOTK-PARTY | .395 | .334 | .362 | .264 | .305 | .332 |
| AMBER | .367 | .313 | .362 | .246 | .294 | .316 |
| BLEU-NRC | .382 | .272 | .322 | .226 | .269 | .294 |
| SENTBLEU-MOSES | .378 | .271 | .300 | .213 | .263 | .285 |
| APAC | .364 | .271 | .288 | .198 | .276 | .279 |
| DISCOTK-LIGHT | .311 | .224 | .238 | .187 | .209 | .234 |
| DISCOTK-LIGHT-KOOL | .005 | .001 | .000 | .002 | .001 | .002 |

**Table 90** Segment-level correlations of automatic evaluation metrics and the official WMT human scores when translating into English

The metric that performed best (0.386) was DiscoTK-Party_tuned, which has proved to work effectively at both system and segment levels. The second position (0.362) in the ranking was occupied by BEER (Stanojević and Sima'an 2014), a metric that does not use linguistic information, except for the distinction between function words and content words, and that combines adequacy features (e.g. matched content words, matched function words), with reordering features, together with a tuning strategy to achieve the best correlation with human judgements (see section 2.2.2.2). The RED family of metrics (Wu et al. 2014) occupies the 3rd (0.356), 4th (0.356), 6th (0.346) and 7th (0.345) positions. This family of metrics uses only the reference dependency tree, which contains lexical and syntactic information and follows a parametric approach.

Information at lexical level includes stems, synonyms, function words and paraphrasing. The different versions of the metric are obtained by following different strategies in the parameters tuning. METEOR is the fifth metric showing a good performance at segment level (0.354). The main novelty in the latest version of METEOR (METEOR Universal) is that it covers previously unsupported languages by using automatically learned linguistic resources (i.e. a function words list and paraphrases), and combining them with a universal parameter for all languages, in opposition to the language-specific parameters used in previous versions. Finally, positions 8$^{th}$ and 9$^{th}$ have gone to UPC-IPA (0.342) and UPC-STOUT (0.338), the latter showing a good result at system level, too. VERTa has not been as effective at segment level as it was at system level. The reason is that linguistic features in each module were based on adequacy and no in-depth study had been conducted on their suitability to evaluate MT quality in general, as already explained. At segment level, the evaluation is more fine-grained than at system-level; thus, a good analysis of intra-module linguistic features would be advisable. Likewise, VERTa's weights were not tuned to correlate well on ranking judgements, as most of the best-performing metrics were. Revisiting linguistic features and how they should be combined to evaluate MT quality, as well as tuning weights to correlate well with ranking judgements, should lead to a better performance of the metric.

### 8.4 Summing Up

In this chapter, a meta-evaluation of VERTa has been carried out. The metric has been evaluated as regards adequacy and fluency, and its participation in the WMT14 metrics shared task has also been presented and analysed. Both quantitative and qualitative analysis have shown that VERTa outperforms other well-known and widely used metrics (i.e. BLEU, METEOR) as well as other linguistically-based metrics (i.e. syntax-based metrics, dependency-based metrics and semantic-based metrics) in both adequacy and fluency evaluations.

As regards adequacy, VERTa's holistic approach and the interaction among its different modules have proved crucial. Results have confirmed that combining information at different linguistic levels leads to better quantitative and qualitative evaluations. This has been proved, not only by VERTa but also by ULC-Combination 1, although the fact that VERTa's modules are less in number and that it focuses on those key linguistic

elements and weights them depending on the evaluation, have been relevant factors for VERTa outperforming ULC-Combination 1. The importance of the Dependency Module must also be highlighted since its flexible matches and language-dependent rules have been key elements to account for different syntactic structures conveying the same meaning. In addition, we also consider that the previous analysis conducted in order to select the dependency parser has contributed to improve the metric's performance.

Regarding fluency, results obtained when correlating with human judgements were worse than those obtained in adequacy. This was not unexpected since similar results were achieved during the development of the metric. However, it must be highlighted that VERTa still outperforms those metrics that were thought to be more fluency-oriented; thus the collaborative approach in VERTa, using different modules that interact with each other (namely the Dependency, LM, N-gram and Morphological Modules) has been essential, once again. In the meta-evaluation on fluency, those metrics dealing with word order, as well as the N-gram Module in VERTa have been quite useful to account for the grammaticality of a segment. In addition, adding the LM Module has also proved useful to match those chunks that do not appear in the reference translations, thus broadening the coverage of lexical items.

On the other hand, we must also mention one of VERTa's weaknesses, the too restrictive performance of the Exact match in the Dependency Module. Since this module only uses the Exact match, VERTa is sometimes too strict in its scores. Therefore, it correlates very well when human judgements assign low scores but it has problems when human judgements are higher. This weakness has tried to be addressed using lexical semantic relations and language-dependent rules in the Dependency Module, but emphasis has to be placed now on widening the coverage of lexical semantic relations.

Once the meta-evaluation of adequacy and fluency was performed we wanted to test VERTa in an official competition, the WMT14 Metrics Shared Task[50], one of the most prestigious venues for research in Machine Translation. Participating in this shared task

[50] http://www.statmt.org/wmt14/metrics-task/

was an interesting challenge because VERTa has been developed to evaluate adequacy and fluency separated, and the metric provides segment scores. However, WMT14 metrics shared task is far from the former types of evaluation, since it seeks for MT quality as a whole and human judgements are based on sentence ranking. Taking all this into account, as well as the lack of time and computational resources derived from the need of the WMT14 Shared Task, and especially since this was the first time that VERTa participated in this type of competition, we must say that the performance of the metrics made the grade. A couple of versions were submitted, VERTa-EQ, same weights assigned to all modules, and VERTA-W, different weights assigned to modules relying on the experiments on adequacy and fluency performed. Results achieved by VERTa were especially noticeable at system level, where both versions achieved the 5$^{th}$ and 6$^{th}$ positions out of 22 metrics and scored above average in all language pairs. At segment level VERTa's performance was worse – both versions achieved 10$^{th}$ and 11$^{th}$ positions out of 18 –, although they scored above average in all language pairs. Still, VERTa and most of the best performing metrics have proved that linguistic information is vital when evaluating MT quality and that combining information at different levels is better than working on partial aspects of language. It must also be emphasized that, in opposition to VERTa, most of the higher-performing metrics were tuned to correlate well with ranking judgements, which definitely boosted their performance. For next editions we would like to rethink those linguistic features suitable to evaluate MT quality; conduct a detailed analysis of intra- and inter-module features; and no doubt, we would also like to tune our metric's weights according to ranking judgements.

This chapter offered a meta-evaluation of VERTa using MT output in English. Our next challenge was porting the English version of VERTa to another language to test whether the metric was easy to port and if the linguistic information used had to be modified and adapted. To this aim, we have ported VERTa to Spanish and we have focused on the evaluation of adequacy. This adaptation is described in the next chapter.

# Chapter 9. Porting VERTa to Evaluate Adequacy for Spanish

Most MT metrics have been developed to evaluate translation into English, since this is the widest spread language in the MT community. However, in the last years several MT evaluation campaigns have been carried out (WMT09-WMT14[51]) boosting the development of MT evaluation metrics not only for English but also for other languages (e.g. Spanish). Some of the metrics participating in these campaigns use lightweight linguistic information at a very specific level or no linguistic information at all (e.g., METEOR (Denkowski and Lavie 2014); AMBER (Chen et al. 2012); TerrorCat (Fishel et al. 2012); TESLA family of metrics (Dahlmeier et al. 2011); WMPF and MPF (Popovic 2011); ROSE (Song and Cohn 2011); ATEC (Wong and Kit 2010); BEER (Stanojević and Sima'an 2014); BLEU (Papineni et al. 2001)), whereas others use a wide range of metrics such as IPA and STOUT (González et al. 2014b). After developing VERTa for English, a new experiment was conducted to test the multilingual capability of this metric. It was our aim to check whether VERTa, which uses richer linguistic knowledge than previously mentioned metrics, could be easily adapted to another language than English, such as Spanish, and if the results obtained outperformed those of other well-known metrics.

This chapter reports the experiments conducted to adapt the English version of VERTa to Spanish in order to evaluate adequacy and is organised as follows: section 9.1 presents the goal of the experiments carried out and describes the data used; section 9.2 describes experiments performed; section 9.3 discusses the differences between Spanish VERTa and English VERTa; section 9.4, compares VERTa to other well-known metrics; and finally, section 9.5 draws some conclusions.

### 9.1 Goals and Data

The experiments conducted in this chapter aim at a) studying which linguistic features were the most appropriate to evaluate the adequacy of a segment in Spanish; b)

[51] http://www.statmt.org/

exploring and finding the most effective combination of VERTa's modules[52] to evaluate adequacy in Spanish output; c) comparing the linguistic information used to evaluate Spanish and English; and finally d) comparing VERTa to other well-known metrics.

In order to perform these experiments part of a corpus developed in the KNOW-2[53] project was used (see section 3.2.1.5). The data contains: 187 WordNet glosses that had been translated from English into Spanish by means of two different systems (Apertium[54] and Google Translator[55]) so as to obtain the hypothesis, four reference translations and human judgements provided by two different judges. Besides, several NLP resources and tools were used to parse the data, as explained in section 3.2.2.2: WordNet was used to get information regarding synonyms, hypernyms and hyponyms, and Freeling's PoS tagger and dependency modules were employed to PoS tag the corpus and obtain its dependency analysis. Experiments were performed at segment level and correlation with human judgements on adequacy was calculated by means of Pearson correlation.

## 9.2 Experiments

Experiments have been carried out to test if VERTa could be easily ported to Spanish in order to evaluate adequacy. To this aim, first those linguistic features that were more suitable to deal with Spanish were selected (section 9.2.1), later, VERTa's modules were combined in order to obtain the best way to correlate with human judgements on adequacy (section 9.2.2).

### 9.2.1 Influence of Linguistic Features

The aim of the first experiment was studying the influence of the linguistic features used in each module.

Regarding the Lexical Module, the same features available in English have also been used in Spanish, with the exception of the partial lemma (see Table 91).

[52] The Semantic and LM Modules are not available for Spanish.
[53] http://ixa.si.ehu.es/know2
[54] http://www.apertium.org/
[55] http://translate.google.com/

| W | Match | Examples | |
|---|---|---|---|
| | | Hypothesis | Reference |
| 1 | Word-form | plantas (*plants*) | plantas (*plants*) |
| 1 | Lemma | era_SER (*was_BE*, imperfect) | fue_SER (*was_BE*, preterite) |
| 1 | Synonymy | prisión (*prison*) | cárcel (*jail*) |
| 1 | Hypernynm | embarcación (*boat*) | barca (*rowboat*) |
| 1 | Hyponym | barca (*rowboat*) | embarcación (*boat*) |

**Table 91** Lexical matches and examples in the Lexical Module

The reason why the partial lemma feature was disregarded is due to the Spanish wider variety of spelling changes in words belonging to the same family, which does not allow for a correct use of this match. This linguistic decision has been confirmed by the correlation obtained when this feature is included in the Lexical Module, which slightly decreases its performance when both reference 1 and all 4 references are available (see Table 93). In addition, the use of hypernyms and hyponyms also seems to improve the performance of the Lexical Module. However, this increase is just a tendency and more data would be needed in order to confirm the appropriateness of such a feature. As for weights, each type of match was assigned the same weight since we did not have enough data to test this type of information.

As regards the Morphological Module, the same matches as in English VERTa were used (see Table 92). Similar to the Lexical Module, all matches were assigned the same weight due to the small amount of data available.

As for the Dependency Module, those matches available in the English VERTa were also used in the Spanish version, except for the No_mod match, which does not correlate well with human judgements when reference 1 is used (see Table 93). This tendency was also confirmed when the 4 references were used: the omission of the No_mod match has a strong positive impact on the correlation of this Module. Regarding the rest of matches, they were assigned the same weight once again due to the shortage of data to carry out further tests. In addition, following the same pattern as in the English version, dependency relations have been assigned a different weight, thus allowing us to distinguish between those relations which are considered more informative (e.g. subject-verb) and those less informative (e.g. determiner-noun). After

testing different weights, the most informative relations have been assigned 1, whereas the least informative ones have been assigned 0.5.

| W | Match | Example | |
|---|---|---|---|
| | | Hypothesis | Reference |
| 1 | Word-form, PoS | plantas, NCFP000 | plantas NFCP000 |
| 1 | Lemma, PoS | era_(SER, VSII1S0) | era_(SER, VSII1S0) |
| 1 | Syn., PoS | prisión, NCFS000 | cárcel, NCFS000 |
| 1 | Hypernym, PoS | embarcación, NCFS000 | barca, NCFS000 |
| 1 | Hyponym, PoS | barca, NCFS000 | embarcación, NCFS000 |

**Table 92** Morphological pairs of matches and examples in the Morphological Module

| Module | Features Changed | Ref. 1 | 4 Refs. |
|---|---|---|---|
| **Lexical** | Partial lemma | 0.4938 | 0.6066 |
| | No partial lemma | **0.5019** | **0.6376** |
| | Hyper./Hypo. | **0.4938** | **0.6376** |
| | No Hyper./Hypo. | 0.4913 | 0.6355 |
| **Morph.** | NO CHANGES | **0.4723** | **0.6007** |
| **Depend.** | No-mod match | 0.4306 | 0.5061 |
| | No No-mod match | **0.4588** | **0.6240** |
| | Dep. relations same weight | 0.4306 | 0.5933 |
| | Dep. relations different weight | **0.4409** | **0.6240** |
| **N-gram** | 2gram-length | **0.3925** | **0.6285** |
| | Sentence-length | 0.3697 | 0.5384 |

**Table 93** Influence of linguistic features

Finally, following the English version of the metric, the N-gram Module was computed over lexical items and had a better performance when 2-gram length was used than when sentence-length grams were used, showing that shorter n-grams length seems to be more appropriate when evaluating adequacy also in Spanish.

### 9.2.2 Combination of Modules

Once the linguistic features for each module were analysed and set, our next step was to explore the combination of such features by combining VERTa's different modules. Table 94 shows the results of each module separately for experiments with one and four references, respectively. In both cases, the module that shows the best correlation is the Lexical Module (0.6376), thus confirming the undeniable fact that lexical semantics plays a key role when evaluating adequacy. The Dependency Module also obtains similar correlations in both cases and occupies the third position in the ranking (0.6240). This indicates that the flexibility of the Dependency Module accounts for syntactically different structures expressing the same meaning.

However, the Morphological and N-grams Module swap positions. The Morphological Module has a significant influence when only one reference is available, since it obtains the second best performance (0.4723). On the other hand, the N-gram Module gets a really low correlation (0.3925). However, when 4 references are used, the second position is occupied by the N-gram Module (0.6285), whereas the Morphological Module seems to be the least influential one (0.6007). It must be noticed, though, that when 4 references are used, the performance of each module is closer in terms of correlation with human judgements than when just reference 1 is considered, as shown in Table 94. Having four different segments against which to compare increases the probabilities of finding a match and reflects the rich reality of language.

| Module | Reference 1 | 4 References |
|---|---|---|
| Lexical Module | **0.5019** | **0.6376** |
| Morphological. Module | 0.4723 | 0.6007 |
| Dependency Module | 0.4588 | 0.6240 |
| N-gram Module | 0.3925 | 0.6285 |

**Table 94** Correlations with human judgements per module, using ref. 1/ using 4 refs.

Furthermore, a thorough analysis of the data shows that the first reference used contains rather free translations, whereas the style of the other three references is closer to the hypothesis. An example of this different style is illustrated in the example below, where references 2, 3 and 4 are clearly closer to the hypothesis than reference 1.

Example 131

**SOURCE**: *the departure of a vessel from a port*

**HYP**: *La salida de un barco de un puerto.*

**REF1**: *Acción de zarpar una embarcación*

**REF2**: *La partida de un navío de un puerto*

**REF3***: La partida de un barco desde un puerto*

**REF4***: La partida de un barco del puerto*

Since VERTa uses similarity measures, it is clear that the preference when selecting a reference to compare the hypothesis with the 4 references available will be reference 2, 3 or 4 which are closer in style than reference 1. This also explains the increase in the performance of the N-gram Module when 4 references are available (from 0.3925 when reference 1 is used to 0.6285, when 4 references are used). The N-gram Module is based on the matches established by the Lexical Module, thus, once lexical matches are set, the n-gram similarity between the hypothesis and reference 2, 3 and 4 is closer than between the hypothesis and reference 1. In order to confirm this point, separate correlations were calculated for each reference. Table 95 shows that for each reference the Lexical Module correlates better with human judgements than the rest of modules, highlighting again the importance of lexical semantics. The module that correlates worst with human judgements is the Morphological Module, except for reference 1, where the N-gram Module is the one that correlates the worst. As explained above, this is mainly due to the free translations in reference 1. In addition, the low correlation of the Morphological Module in most of the references was expected, as this module seems more appropriate to deal with fluency issues. As regards the use of the Dependency Module, it proves effective in most of the references.

| Reference | Module | Pearson Correlation |
|---|---|---|
| 1 | Lexical Module | **0.5019** |
| | Morphological Module | 0.4723 |
| | Dependency Module | 0.4588 |
| | N-gram Module | 0.3925 |
| 2 | Lexical Module | **0.5736** |
| | Morphological Module | 0.5020 |
| | Dependency Module | 0.5619 |
| | N-gram Module | 0.5484 |
| 3 | Lexical Module | **0.5224** |
| | Morphological Module | 0.4744 |
| | Dependency Module | 0.5087 |
| | N-gram Module | 0.5081 |
| 4 | Lexical Module | **0.4779** |
| | Morphological Module | 0.3893 |
| | Dependency Module | 0.4451 |
| | N-gram Module | 0.4470 |

**Table 95** Pearson correlation per module using each reference separately

In order to make a final decision on the combination of VERTa's modules, all references were used. From a linguistic point of view, and taking into account the type of evaluation and the characteristics of the language evaluated, those modules that seem to be the most appropriate were first the Lexical and Dependency Module, coinciding with the conclusions reached in the English version. The Lexical Module accounts for semantics at word level because it uses synonymy and hypernymy/hyponymy relations. In addition, it must be noticed that dependency relations are an interface between syntax and semantics since they account for the internal relations in a sentence, moving away from its surface structure. Hence, the Dependency Module is a good candidate to evaluate sentence semantics, which has already been proved for English adequacy (see section 6.7). As for the N-gram and Morphological Modules, the N-gram Module does not seem to play a key role when evaluating adequacy, although it is more important than the Morphological Module, since word order in a sentence has a stronger influence

on meaning than inflectional morphology. Bearing all this in mind, modules' weights were first assigned manually following both linguistic criteria and weights obtained for the English data (see Table 96). It must be noticed that in the English version the Morphological Module was disregarded when evaluating adequacy (see section 6.7); however, in Spanish, we consider that it must be taken into account because of its richer inflectional morphology. Later, in addition, in order to calculate an upper-bound for the weight tuning, all possible weight combinations were tuned automatically using a 0.01 step. The results obtained (see Table 96) confirmed our initial hypothesis that the highest weights should be assigned to the Lexical Module and the Dependency Module as they account for the meaning of the sentence, whereas the N-gram Module and especially the Morphological Module play a minor role when assessing adequacy in Spanish.

| | **Manual Weight** | **Automatic Weight** |
|---|---|---|
| **Lexical Module** | 0.45 | 0.46 |
| **Morphological Module** | 0.05 | 0.03 |
| **Dependency Module** | 0.40 | 0.32 |
| **N-gram Module** | 0.10 | 0.19 |
| **PEARSON CORREL.** | **0.6596** | **0.6611** |

**Table 96** Correlations obtained when using manual and automatically tuned weights

Now that the linguistic features that help in testing adequacy have been explored and discussed, our next goal was to compare the English version of VERTa to the Spanish version of VERTa, as well as their corresponding linguistic features.

### 9.3 Spanish VERTa vs. English VERTa

It was also of our interest to compare VERTa's performance when evaluating Spanish and its performance when evaluating English data. Results obtained for Spanish contrast with those obtained in Chapter 6 for English (see Table 97).

| | English VERTa | Spanish VERTa |
|---|---|---|
| **Lexical Module** | 0.47 | 0.46 |
| **Morphological Module** | 0 | 0.03 |
| **Dependency Module** | 0.43 | 0.32 |
| **N-gram Module** | 0.05 | 0.19 |
| **Semantic Modules**[56] | 0.05 | |
| **PEARSON CORREL.** | **0.781** | **0.661** |

**Table 97** VERTa's correlation for English data

Although it is difficult to compare the data set used for Spanish and the one used for English, because their size, genres and style are very different, some preliminary conclusions can be drawn.

As regards intra-module linguistic features, firstly, due to the wider variety of spelling changes in Spanish, the partial lemma feature, which has proved suitable in English, was disregarded in Spanish; secondly, the same features as in English have been kept in the Morphological Module; thirdly, similar to the English version, different weights were assigned to dependency labels in the Dependency Module; finally, also following the thread of the English version, short n-grams calculated over lexical matches were preferred in the N-gram Module to evaluate adequacy.

Regarding the combination of modules, firstly, the Lexical and Dependency Modules are the most effective and appropriate ones to evaluate the adequacy of a segment both in English and Spanish; secondly, the N-gram Module should also be used but its influence on determining the adequacy of a segment is not crucial; finally, automatically tuned weights confirmed that whereas in English the Morphological Module does not prove effective to evaluate adequacy, in Spanish it might be taken into account, although its role is less significant than the Lexical and Dependency Module's. The reason why this module should be slightly considered in Spanish but not in English is that Spanish shows a richer inflectional morphology than English, although its influence would probably be stronger if fluency was assessed. It must also be noticed that the N-gram Module is assigned a higher weight for Spanish than for English, which is a bit

[56] The Semantic Module is not available for Spanish yet.

contradictory since English shows a stricter word order than Spanish. This might be explained due to the different segment length of the Spanish and English corpus. The segments in the Spanish data are dictionary definitions, thus they are shorter than the English segments, which belong to news and contain complex structures as well as a frequent use of subordination. The shortness of the Spanish segments results in a greater importance of the word order, even if adequacy is evaluated.

### 9.4 Comparing VERTa to other MT Metrics

Once experiments aimed at analysing the suitability of linguistic features to evaluate adequacy in Spanish were conducted and discussed, the most natural step was to compare VERTa to other well-known metrics. Metrics used to compare VERTa were BLEU, METEOR-ex (only exact matching), METEOR-st (exact matching plus stemming) and METEOR-pa (exact matching, stemming and paraphrasing) and a set of linguistically-based metrics available in the Asiya tool (Giménez and Márquez 2010a; González et al. 2014). In this set of metrics, a couple of them use shallow parsing (SP-Op(*) and SP-Oc(*)); others capture similarities between dependency trees (DPm-Ol(*), DPm-Oc(*) and DPm-Or(*)); finally, others compare similarities between constituent parse trees (CP-Op(*)and CP-Oc(*)) (refer to section 2.2.2.4 for further details). Results obtained are shown in Table 98.

Results obtained show that VERTa outperforms the rest of metrics (0.6611), although the METEOR family also obtains good results, especially the version that uses paraphrasing (0.6212). This indicates that when assessing adequacy the metric must be flexible enough to account for lexical semantic relations and different ways to express the same meaning. N-gram-based metrics, such as BLEU, do not show a good correlation with human judgements (0.5551), mainly because they are too rigid and account for word order, as a consequence, the omission of a single determiner is penalised. Other linguistically-based metrics show a lower performance than VERTa, this is mainly due to the fact that they do not use any kind of information regarding lexical semantics, thus showing a lower flexibility than VERTa or METEOR. It is also noticeable the lower performance of the metric that uses information on dependency relations (DPm-Or(*)) (0.4483), which was expected to obtain a higher correlation with

human judgements. Such a low performance might be due to the performance of the parser used for Spanish.

| Metric | Pearson Correlation |
|---|---|
| VERTa | **0.6611** |
| METEOR-ex | 0.6017 |
| METEOR-st | 0.6152 |
| METEOR-pa | 0.6212 |
| BLEU | 0.5551 |
| SP-Op(*) | 0.5770 |
| SP-Oc(*) | 0.5624 |
| DPm-Ol(*) | 0.4285 |
| DPm-Oc(*) | 0.5616 |
| DPm-Or(*) | 0.4483 |
| CP-Op(*) | 0.5246 |
| CP-Oc(*) | 0.5684 |

**Table 98** Comparison between VERTa and other well-known metrics

Correlations aside, data was also analysed in detail in order to compare VERTa's and METEOR-pa's performance. This analysis indicates that synonymy relations and the Dependency Module play a key role when comparing both metrics and are the main reason why VERTa outperforms METEOR-pa, as illustrated by Examples 132 and 133.

Despite not being a very natural sentence, the hypothesis segment in Example 132 conveys the meaning of the source segment. Synonymy helps in matching *deberes* ("duties") and *tareas* ("tasks"), as well as *criado* ("manservant") and *sirviente* ("servant").

Example 132

**SOURCE**: *the performance of duties by a waiter or servant; "that restaurant has excellent service "*

**HYP**: *El rendimiento de deberes por un camarero o criado; "aquel restaurante tiene servicio excelente".*

**REF**: *Cumplimiento de la tarea de un camarero o un sirviente; "este restaurante tiene un servicio excelente"*

In the example below, the hypothesis segment communicates the meaning of the source segment, although it is slightly disfluent. In addition, the reference translation is rather free, since *en el intento de* ("in the attempt to") has been added despite the fact that it does not appear in the source sentence. Fortunately, the Dependency Module helps in maintaining the core meaning of the sentence and accounts for the relation of *fracaso* ("failure") and *mantener* ("maintain") despite the addition of *en el intento* ("in the attempt to").

Example 133

**SOURCE**: *a failure to maintain a higher state*

**HYP**: *Un fracaso de mantener un estado más alto.*

**REF**: *Fracaso en el intento de mantener un estado superior*

This section has compared VERTa's performance to that of other well-known metrics and has shown that VERTa outperforms them. Next, the main findings of this chapter aimed at porting VERTa to Spanish to evaluate adequacy are summarized.

### 9.5. Findings

Experiments indicate that VERTa can be easily adapted to other languages than English. The effort behind this adaptation to Spanish could be quantified in terms of the tasks defined in sections 9.1 and 9.2. Firstly, dealing with different linguistic phenomena that are not present in English, such as the wider variety of spelling changes or a richer inflectional morphology; and secondly, changing the NLP tools and resources used in English for those relevant to Spanish. In addition, despite the fact that the existence and quality of the different NLP analyzers for languages other than English could be an issue, this does not seem to be the case for Spanish or, at least, it does not seem to affect VERTa's performance.

Experiments have also shown that when evaluating adequacy for both Spanish and English the Lexical and Dependency Modules are the most effective ones, followed by

the N-gram Module. Actually, in the N-gram Module, short n-grams over lexical items are preferred to evaluate adequacy in both languages. On the other hand, due to language particularities the Morphological Module should be used when evaluating the adequacy of Spanish MT output and disregarded when evaluating English.

It has also been proved that VERTa gets better results than other well-known metrics, leading to the conclusion that a more collaborative approach that accounts for different aspects of language achieves a better correlation with human judgements than those approaches that focus on more partial aspects. Even when the reference translations are rather free VERTa's results are better, mainly due to the help of the Dependency Module and lexical semantics; in other words, thanks to the use of a more collaborative approach.

# Chapter 10. Main Contributions and Future Work

This final chapter begins revisiting the hypotheses initially formulated (section 10.1) to check whether they have been corroborated by our research and then moves on to a summary of this thesis' major contributions (section 10.2). The present chapter finishes pointing out new research directions on the use of linguistic features to evaluate MT output and new applications of the MT metric developed (section 10.3).

## 10.1 Revisiting our Initial Hypotheses

In the introductory chapter, four different hypothesis and their respective sub-hypotheses were formulated (please refer to Chapter 1 for a full description) as the basis of this research. The following aims at revisiting these hypotheses and checking whether they have been confirmed.

**Hypothesis 1.** *A linguistic analysis can help to clarify what linguistic features should be used and how they should be combined to evaluate MT output.*

This hypothesis has been answered by means of the linguistic analysis of the data conducted (see Chapter 4) and the qualitative analysis performed during the experiments for English (see Chapters 6 and 7) and to port VERTa to Spanish (Chapter 9). Although usually the correlations with human judgements and the qualitative analysis coincide, we have found evidence that correlations with human judgements are rather superficial and that a deep, thorough qualitative analysis can help in taking better decisions as regards which linguistic features should be used, thus providing more trustful results. This was especially evidenced when checking linguistic features to evaluate fluency, in particular with the use of synonyms (see section 7.2.1) and the No_mod match (see section 7.4.1.2). Although correlations with human judgements advised against the use of synonyms, linguistic analysis proved that the decrease in the correlation was not related to any issues regarding the grammaticality of the sentence. Furthermore, correlations with human judgements also advocated for the use of the No_mod match since adding this type of match resulted in an increase in the correlation.

On the other hand, the linguistic analysis performed showed that the benefits of using that match were only related to the widening of the lexical coverage and, on the contrary, its use involved accepting ungrammatical syntactic structures, which directly affected the fluency of a segment.

As for Spanish, following our linguistic criteria and due to its rich inflectional morphology, we considered that the Morphological Module could not be disregarded, even if only adequacy was tested. Later on, experiments confirmed this decision showing that the performance of the metric improved if the Morphological Module was used.

Therefore, according to our experience, the combination of correlations with a more qualitative analysis is recommended in order to obtain a more reliable and complete evaluation of the suitability and influence of linguistic knowledge in MT evaluation.

**Hypothesis 2**: *Linguistic features would be more or less appropriate depending on the type of evaluation, either adequacy or fluency.* This can be broken down into the following specific points:

i. *Organising linguistic information at different levels and aiming at different tasks might help to detect MT errors, which might be especially useful to improve knowledge-based MT systems.*

   Dividing and organizing the MT metric into different levels helps in detecting MT errors easily. Since each module in VERTa can work individually, partial aspects of language can be checked: the Morphological Module can help to identify lack of agreement between the subject and the verb (section 7.3.1); extracting No_mod matches in the Dependency Module can help to identify ungrammatical chunks due to untranslated words or bad translations affecting mainly function words (i.e. prepositions, determiners, pronouns and conjunctions) (section 7.4.1.2), whereas extracting No_head matches can identify ungrammatical chunks due to mistranslations or untranslated content words (e.g. nouns and verbs) (section 7.4.1.3).

ii. *Lexical semantics helps to evaluate adequacy. Most linguistically-enhanced metrics use synonyms, but we think that using other type of lexical semantic relations, such as hypernyms and hyponyms might help to evaluate adequacy.*

Our experiments have proved that hypernymy and hyponymy relations cannot be entirely disregarded when evaluating MT output. Although final conclusions cannot be reached yet, since a larger amount of data would be necessary, experiments have shown that, at some point, these semantic relations might be helpful when one reference is available (see section 6.2.2). Interestingly enough, this feature has also proved effective for fluency in English (see section 7.2.1) and also when porting our metric to Spanish (see section 9.2.1). This indicates that the use of hypernymy and hyponymy as a new feature to evaluate MT is still an open question which is worth analysing in depth.

iii. *Depending on how syntactic information is used it can help to evaluate adequacy or fluency.*

Although initially syntactic information seems more suitable to evaluate fluency, and this has been the goal of most of the syntax-based metrics, our experiments have proved that dependency relations can also be effective to evaluate adequacy. The Dependency Module is suitable for evaluating adequacy when more flexible types of matches are allowed since they account for syntactically different structures conveying the same meaning or for not totally grammatical structures that can still be understood (see section 6.4). Moreover, in our experiments, the Dependency Module has proved more effective than other SR-based metrics to evaluate adequacy (see section 8.1.1) at segment level, in opposition to what Lo and Wu (2010) stated. This indicates that dependency relations can also account for sentence semantics since they account for the compositionality of the sentence.

On the other hand, if the Dependency Module is used in a more restrictive way – using only the Exact match – and dependency relations are distributed and weighed according to the position they occupy inside the dependency tree, the Dependency Module can also be useful to evaluate the fluency of a segment.

Dependency relations are somewhere between syntax and semantics, thus depending on how they are used, they can account for either adequacy or fluency.

iv. *Information regarding Semantic Roles and Named Entities has been used to evaluate adequacy (Lo et al. 2012). We also think that other semantic information such as Sentiment analysis, NE linking and identification of Time Expressions can help to evaluate adequacy.*

Information regarding NEL, identification of Time Expressions and Sentiment analysis has been tested to evaluate adequacy in the Semantic Module. Unfortunately, our experiments have shown that none of these features helps to evaluate adequacy individually because they evaluate only partial aspects of the segment, however, they have proved effective in combination with other linguistic information (i.e. dependency relations and lexical semantics). Thus, indicating that they cannot be disregarded when evaluating adequacy.

**Hypothesis 3:** *We think that working on different evaluation tasks might not only be useful to identify which linguistic features are the most appropriate depending on the type of evaluation (adequacy or fluency) but also how they should be combined.*

i. *In order to evaluate the fluency of a segment, that information aimed at checking the grammaticality of a sentence seems to be the most convenient: morphosyntactic information (i.e. lemma, PoS), word order and dependency relations.*

In our experiments, both correlations with human judgements and the qualitative analysis have proved that those linguistic features with a strong influence on the evaluation of fluency are dependency relations and n-grams calculated over PoS (section 7.8). This information looks after well-constructed structures at phrase and clause level, as well as word order, which is especially important since English shows a quite fixed word order. Likewise, n-grams over PoS relate inflectional morphology and syntax accounting for morphosyntactic features such as subject-verb agreement. Information related to PoS in isolation (the Morphological Module) also contributes but in a noticeably smaller way. This is

related to the fact that English does not show a rich inflectional morphology and the n-grams are already computed over PoS tags.

ii. *In order to evaluate the adequacy of a segment, that information related to both lexical and dependency relations seems to be the most relevant information. According to the principle of compositionality (Frege's Principle) "the meaning of a whole is a function of the meaning of the parts and of the way they are syntactically combined". Thus, the interaction between lexical semantics and dependency relations should account for the meaning of the sentence.*

The linguistic features that play a major role in the evaluation of the adequacy of a segment are definitely those included in the Lexical Module and in the Dependency Module, as confirmed by the experiments in section 6.7. Those in the Lexical Module – word-form, synonyms, lemma and partial lemma – deal with the meaning of lexical items, whereas those related to the Dependency Module – also including information about word-form, synonyms, lemma and partial lemma – account for how the sentence is built. In addition, linguistic features included in the Semantic Module also contribute slightly to evaluate the adequacy of a segment. Their contribution is small because they account for very partial aspects of the meaning of a sentence, however, they cannot be disregarded. On the other hand, we did not expect that information regarding word order would have any influence since dependency relations were already used, however, according to the experiments carried out, the N-gram Module contributes slightly to the evaluation of adequacy, indicating that word order cannot be entirely disregarded either.

**Hypothesis 4**: *Depending on the source and target language, the type of linguistic features used and how they are combined might vary. Thus, porting a linguistically-enhanced MT metric to a new language may involve considering its linguistic characteristics and reflecting them on how linguistic features are used in the metric.*

To confirm this hypothesis we have ported our MT metric, VERTa, from English into Spanish to evaluate adequacy. Porting the MT metric to Spanish was not hard work, the adaptation was quite straightforward. We adapted most of the modules, except for the

Semantic and LM modules, we changed the NLP tools to parse English for others to parse Spanish and we have used WordNet 3.0 in Spanish. As for those characteristics related to the source and target languages, we had to take into account some language-dependent characteristics such as the wider variety in Spanish spelling which was reflected in the omission of the partial lemma, and the wider range of different verb tenses in Spanish which poses problems when translating from English. This confirms our general hypothesis that when porting an MT metric, both source and target languages have to be considered.

i. *Information on PoS might be disregarded when evaluating adequacy in English, but it might be useful when addressing Spanish.*

   In our experiments with the English version of VERTa, PoS information did not prove effective to address adequacy in English (see section 6.3), but according to some preliminary experiments performed, it slightly contributes to the evaluation of adequacy in Spanish (see section 9.2.2). The fact that this information is used in Spanish whereas in English is disregarded, responds to the richer inflectional morphology that Spanish shows when compared to English. On the other hand, its minor contribution to the final combination of linguistic features for Spanish is due to the type of evaluation addressed, especially focused on the meaning of the sentence.

ii. *Word order might have a stronger influence when evaluating English than when evaluating Spanish, since word order in Spanish is more flexible than in English.*

   This hypothesis was not confirmed by the experiments performed (section 9.3). In these experiments the influence of the N-gram Module, accounting for word order, was even stronger when Spanish was evaluated. First, we found it quite shocking because English shows a more fixed word order than Spanish, thus the N-gram Module for these two languages was expected to work in a completely opposite way. However, as stated in section 9.2, the strong influence of the N-gram Module might respond to the rather literal style of the reference translations and the type of sentence structure in the data used, a very fixed

structure since the data used to port VERTa to Spanish comes from WordNet glosses, thus similar to dictionary definitions.

## 10.2 Major Contributions

In the present study we have provided a **qualitative analysis and evaluation of the linguistic features that play an important role in the evaluation of MT output**. The starting point of this empirical study was the linguistic analysis of data to identify which linguistic information would be useful to evaluate MT output and which MT metrics should be considered.

As a result of this study, **we have proposed a classification and description of the information that must be considered when evaluating MT output, including both the MT errors that are detected and positive features**. One of the major characteristics of this classification and description is that it is not based on a particular system as previous works (Vilar et al. 2006; Farrús et al. 2010), but data provided by several systems and different MT technologies has been used. Likewise, it does not only focus on MT errors but it also addresses linguistic phenomena that should not be penalized in an MT evaluation.

Considering the linguistic information identified in our study, **we have developed a linguistically-motivated MT metric, VERTa**. This metric was used as a tool to check the suitability of the linguistic features selected and how they should be combined in order to evaluate adequacy and fluency. Several experiments were conducted on a per-module basis and also in a combination of modules until **we found out which linguistic features should be employed and how they should be combined in an MT metric to evaluate adequacy and fluency in English. The resulting features and their combinations go beyond a quantitative analysis and head towards a more qualitative approach**, thus moving away from combining a wide range of metrics, which makes it difficult to check their contribution to the analysis, and from using machine learning techniques that require a large amount of data. Our analysis to identify and select the linguistic information and how it should be combined has linked traditional correlations with human judgements with a linguistic analysis of the data every time a new linguistic feature was added. The use of correlations has been useful

as a point of departure for our analysis, to refine weights and guide our understanding of the modules in VERTa and their interaction. The linguistic combination to evaluate adequacy that we propose involves mainly information at lexical level (i.e. word-form, synonyms, hypernyms, hyponyms, lemma and partial lemma) and at syntactic level (i.e. dependency relations). Besides, in a lower degree it also requires word-order features (i.e. n-grams) and other semantic-related features (NER, NEL, Time Expressions and Sentiment analysis). The translation of these features into VERTa's modules is the use of the Lexical, Dependency, N-gram and Semantic Modules. As regards fluency, the linguistic combination involves using mainly information regarding dependency relations, word order and PoS features. Actually, it must be highlighted that at this point linguistic features have interacted with an LM, thus combining a reference-based approach with a target-based approach. As regards VERTa's modules, this information corresponds to an important contribution of the Dependency, LM and N-gram Modules and a minor use of the Morphological Module.

During the experiments, state-of-the-art linguistic features were revisited and **we found out that dependency relations, traditionally more fluency-oriented features, can also be used to evaluate adequacy, achieving very good results, indeed**.

In addition, we have also tested the use of unfrequently used features related to textual entailment: NE linking, Time Expressions identification and Sentiment analysis. Although their individual use does not help in the evaluation of MT output, **we have proved that the interaction of NER, NEL, Time Expressions and Sentiment analysis is effective to evaluate adequacy in combination with other adequacy-oriented linguistic features.**

On the other hand, **we have tested linguistic features that had not been used before**. From these features, **our experiments indicate that the use of hypernymy and hyponymy relations should not be entirely disregarded in MT evaluation, especially when only one reference translation is available**. Although no final conclusions can be drawn yet, we consider that this experiment opens the door to the use of more sophisticated lexical semantic relations in MT evaluation.

Finally, our last experiment was porting our MT metric to Spanish, so as to check the degree of difficulty in adapting our metric to a new language and also to perform some experiments which have helped to check the suitability of the linguistic features included in VERTa to evaluate adequacy in Spanish. From this, we can first conclude that porting VERTa, despite using linguistic information, was an easy task and now **there is a Spanish version of VERTa to evaluate adequacy available**.

From this last experiment **we have also concluded that in both English and Spanish, the linguistic information that contributes the most in the evaluation of adequacy is dependency relations and lexical information (i.e. word-form, synonyms, hypernyms, hyponyms, lemma and partial lemma)**. The experiments performed have shown that both the Dependency and Lexical Modules, together with the linguistic features included in each of them, strongly contribute to the evaluation of adequacy in both languages. In addition, English and Spanish data belong to very different genres (see section 3.2.1 for further details on the data used), which indicates that the combination of **the Dependency and the Lexical Module to evaluate adequacy is valid across genres**.

Last but not least, since VERTa is organized in different modules, covering different types of language dimensions, and evaluates fluency and adequacy separately, **our metric can also serve as a tool for error analysis**. Some examples of this use have been provided in Chapter 7.

### 10.3 New Research Directions

The work presented in this thesis is just a small step towards the qualitative analysis of linguistic features in MT evaluation. Actually, there is still a long way to go in order to improve MT metrics from a qualitative perspective. Here we offer a summary of those lines we would like to study further.

Firstly, since our work has been partly inspired by Giménez and Marquez (2010a/b), who used correlations with human judgements and different data sets to find the best combination of linguistic features to evaluate MT quality, we would like to widen the scope of our qualitative analysis by using a larger amount of data and including different data sets. The use of a larger amount of data would allow us to reach more

conclusive weights, especially as regards intra-module weights which, as shown in Chapters 6 and 7, could not be considered as final weights, but just a tendency.

Secondly, we would like to extend our experiments for the Spanish version of VERTa. As described in section 3.2.1.5 and section 9.1, the Spanish data used is rather small and very different from the English data used. Thus we are interested in using a larger data set so that we can reach final conclusions, especially on some controversial points such as avoiding the No_mod match in the Dependency Module and the intra-module weights (section 9.2.1). In addition, the experiments reported for Spanish were only aimed at evaluating the adequacy of a segment, thus, we would be interested in performing experiments to evaluate fluency, which we think would be more language-dependent. Likewise, we have planned to implement the Semantic and LM Modules for the Spanish version of VERTa too.

Thirdly, we are very interested in continuing working on the use of new features in MT evaluation. In this sense, since we could not reach final conclusions on the use of hypernymy and hyponymy relations we would like to use more data so as to provide a wider analysis. In addition, we would also like to refine the use of such semantic features by employing a word sense disambiguation system. We believe that the use of such a system might lead to a better performance of these lexical semantic relations, and thus to a better performance of the metric.

Fourthly, we would like to add Semantic Role information to VERTa. Although VERTa outperformed SR-based metrics in the evaluation performed (see section 8.1.1) we consider that Semantic Roles in combination with lexical semantics and dependency information will provide a better and wider coverage of those features involved in the adequacy of a segment.

Fifthly, some of the features used in the Semantic Module, mainly NEs and Time Expressions, are aimed at matching expressions that contain the same meaning but differ in their form (section 6.6). Actually, several instances of them were found in the linguistic analysis of the data (see section 4.2.1 and 4.2.2). We think that the NEL and Time Expressions metrics could be used in a pre-process stage to identify these expressions conveying the same meaning but differing in their form and substitute them

for a normalized form. This normalization will probably help the NLP tools used for parsing both hypothesis and reference segments, thus probably resulting in a better performance of the metric.

Finally, in the above paragraph, the NLP tools used have been mentioned. During our experiments we found that these tools made mistakes that affected the performance of our metric. Some of these errors that we have already detected are usually caused by the PoS tagger (see 7.3.1), such as not distinguishing proper nouns from upper case words. Although these errors are caused by the PoS tagger, they are propagated through the parsing chain and finally the metric's performance is also affected. Thus, we are particularly interested in detecting parse errors and exploring the impact that they have on the performance of our metric.

On a different note, we would also like to apply VERTa to another NLP task: Recognizing Textual Entailment (RTE). Actually, Textual Entailment (TE) has been used in some MT metrics (Padó et al. 2009; Castillo and Estrella 2012) since somehow, RTE and evaluation of MT using references (at least when evaluating adequacy) are not that far, as both of them compare a hypothesis and reference segment and try to find out if they are semantically similar. We would like to check if VERTa can also be useful in this NLP task.

# References

Akiba, Y., Imamura, K., and Sumita, E. (2001). Using Multiple Edit Distances to Automatically Rank Machine Translation Output. *Proceedings of Machine Translation Summit VIII*, pp. 15-20. Santiago de Compostela, Spain.

Albrecht, J. S. and Hwa, R. (2007a) A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 880-887. Prague, Czech Republic.

Albrecht, J. S. and Hwa, R. (2007b) Regression for sentence-level MT evaluation with pseudo references. In: *ACL 2007 Proceedings of the 45th annual meeting of the Association for Computational Linguistics*, pp. 296-303. Prague, Czech Republic.

Agarwal, A. and Lavie, A. (2008). METEOR, M-BLEU and M-TER: Flexible Matching and Parameter Tuning for High-Correlation with Human Judgments of Machine Translation Quality. *Proceedings of the ACL2008 Workshop on Statistical Machine Translation*. Columbus, Ohio, USA.

Atserias, J., Comelles, E. and Mayor, A. (2005). Txala: un analizador libre de dependencias para el castellano. *Procesamiento del Lenguaje Natural*, (35):455-456, September.

Atserias, J., Blanco, R., Chenlo, J. M. and Rodriguez, C. (2012). *FBM-Yahoo at RepLab 2012. CLEF (Online Working Notes/Labs/Workshop)*. September 20, 2012.

Attardi,G. (2006). Experiments with a Multilanguage Non-Projective Dependency Parser. *Proceedings of the Tenth Conference on Natural Language Learning*, pp. 166-170. New York, USA.

Avramidis, E., Popović, M., Vilar, D. and Burchardt, A. (2011). Evaluate with Confidence Estimation: Machine ranking of translation outputs using grammatical features. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pp 65-70. Edinburgh, Scotland, UK.

Babych, B., and Hartley, T. (2004). Extending the BLEU MT Evaluation Method with Frequency Weightings. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Columbus, Ohio, USA.

Banerjee, S., and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*. Michigan, USA.

Bassnett, S. (1980). *Translation Studies*. Routledge, New York.

Bojar, O., Ercegovčević, M., Popel, M. and Zaidan, O. (2011). A Grain of Salt for the WMT Manual Evaluation. *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 1–11. Edinburgh, Scotland,

Bojar, O. and Wu, D. (2012). Towards a Predicate-Argument Evaluation for MT. *Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 30-38. Jeju, Republic of Korea.

Briscoe, Ted, J. Carroll and R. Watson. (2006). The Second Release of the RASP System. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions,* pp. 77-80. Sydney, Australia.

Callison-Burch, Fordyce, C., Koehn, P, Monz, Ch. and Schroeder J. 2007. (Meta-) Evaluation of Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 136-158. Prague, Czech Republique.

Callison-Burch, Ch., Koehn, P., Monz, Ch., Post, M., Soricut, R. and Specia, L. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. *Proceedings of the Seventh Workshop on Statistical Machine Translation,* pp. 10-51, Montréal, Canada.

Carreras, X., Chao, I., Padró, L., and Padró, M. (2004). FreeLing: An Open-Source Suite of Language Analyzers. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC),* pp. 239–242. Lisbon, Portugal.

Carreras, X., Màrquez, L., and Castro, J. (2005). Filtering-Ranking Perceptron Learning for Partial Parsing. *Machine Learning*, 59, 1–31. Springer. The Netherlands.

Castillo, J. and Cardenas, E. (2010). Using sentence semantic similarity based on WordNet in recognizing textual entailment. *Proceedings of the 12th Ibero-American conference on Advances in artificial intelligence*. Bahía Blanca, Argentina.

Castillo, J. and Estrella, P. (2012). Semantic Textual Similarity for MT Evaluation. *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pp. 52-58. Montréal, Canada.

Chang, Y.S. and Ng., H.T. (2008). MAXSIM: A maximum similarity metric for machine translation evaluation. *Proceedings of the ACL-08: HLT*, pp. 55-62. Columbus, Ohio, USA.

Chang A. X. and Manning, Ch. D. (2012). SUTIME: A Library for Recognizing and Normalizing Time Expressions. *8th International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey.

Charniak, E., and Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*. Michigan, USA.

Chen, B. and Kuhn, R. (2011). AMBER: A Modified BLEU, Enhanced Ranking Metric. *Proceedings of the 6th Workshop on Statistical Machine Translation*, pp. 71-77. Edinburgh, Scotland, UK.

Chen, B., Kuhn, R. and Foster, G. (2012). Improving AMBER, an MT Evaluation Metric. *Proceedings of the 7th Workshop on Statistical Machine Translation*, pp. 59-63. Montréal, Canada.

Church, K. W., and Hovy, E. H. (1993). Good Applications for Crummy Machine Translation. *Machine Translation*, 8(4), pp. 239–258. Springer. The Netherlands.

Ciaramita, M. and Y. Altun. (2006). Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Clark, S., and Curran, J. R. (2004). Parsing the WSJ using CCG and Log-Linear Models. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 104–111. Ohio, USA

Comelles, E., Arranz, V. and Castellon, I. (2010). Constituency and Dependency Parsers Evaluation. SEPLN (ed.), *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Valencia: SEPLN, v. 45, pp. 59-66. Valencia, Spain. ISSN: 1135-5948

Corston-Oliver, S., Gamon, M., and Brockett, C. (2001). A Machine Learning Approach to the Automatic Evaluation of Machine Translation. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL),* pp. 140–147. Toulouse, France.

Dahlmeier, D., Liu, Ch., and Ng, H. T. (2011). TESLA at WMT 2011: Translation Evaluation and Tunable Metric. *Proceedings of the 6th Workshop on Machine Translation*, pp. 78-84. Edinguburgh, Scotland, UK.

De Marneffe, M.C., MacCartney, B. and Manning, C.D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. *Proceedings of the 5th Edition of the International Conference on Language Resources and Evaluation (LREC-2006)*. Genoa, Italy.

Denkowski, M. and Lavie, A. (2010). METEOR-NEXT and METEOR Praraphrase Tables: Improved Evaluation Support for Five Target Languages. *Proceedings of the Joint 50th Workshop on Statistical Machine Translation and MetricsMATR*, pp. 339-342, Uppsala, Sweeden.

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. *Proceedings of the 6th Workshop on Statistical Machine Translation*, pp. 85-91. Edinburgh, Scotland, UK.

Denkowski, M. and Lavie, A. (2014). Meteor Universal. Language Specific Translation Evaluation for Any Target Language. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376-380. Baltimore, Maryland USA.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *Proceedings of the 2nd International Conference on Human Language Technology,* pp. 138–145. San Diego, California.

Estrella, P., Popescu-Belis A. and Underwood N. (2005). Finding the System that Suits you Best: Towards the Normalization of MT Evaluation. *Proceedings of the 27th International Conference on Translating and the Computer, ASLIB*. London, UK.

Farrús, M., Costa-Jussà, M. R., Mariño, J.B. and Fonollosa, J. A. R. (2010). Linguistic-based Evaluation Criteria to identify Statistical Machine Translation Errors. *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*. Saint Raphael, France.

Fellbaum, Ch. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fillmore, Ch. (1968). The Case for the Case. In Bach and Harms (Ed.): *Universals in Linguistic Theory*. New York: Holt, Rinehart and Witson.

Fishel, F., Sennrich, R., Popović, M., Bojar, O. (2012). TerrorCat: a Translation Error Categorization-based MT Quality Metric. *Proceedings of the 7th Workshop on Statistical Machine Translation*, pp. 64-70. Montréal, Canada.

Gamon, M., Aue, A., and Smets, M. (2005). Sentence-Level MT evaluation without reference translations: beyond language modeling. *Proceedings of EAMT05*, pp. 103–111. Budapest, Hungary.

Gautam, S. and Bhattacharyya, P. (2014). Layered: Metric for Machine Translation Evaluation. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 387-393. Baltimore, Maryland, USA.

Giménez, J. and Márquez, L. (2004). SVM Tool: A general POS tagger generator base don Support Vector Machines. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, vol. I, pp. 43-46. Lisbon, Portugal.

Giménez, J. 2008a. *Empirical Machine Translation and its Evaluation*. PhD Thesis. Universitat Politècnica de Catalunya. Spain.

Giménez, J. and Márquez, Ll. (2008b). Discriminative Phrase Selection for Statistical Machine Translation. In C. Goutte, N. Cancedda, M. Dymetman and G. Foster (eds.) *Learning Machine Translation*. NIPS Workshop Series. MIT Press.

Giménez, J. and Márquez, Ll. (2010a). "Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation." *The Prague Bulletin of Mathematical Linguistics*, No. 94, pages 77-86.

Giménez, J. and Márquez, Ll. (2010b). Linguistic Measures for Automatic Machine Translation Evaluation. *Machine Translation*, 24(3–4),77–86. Springer. The Netherlands.

González, M. and Giménez, J. (2014). *Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. Technical Manual 3*.0. Universitat Politècnica de Catalunya. Spain.

González, M., Barrón-Cedeño A., Márquez, Ll. (2014). IPA and STOUT: Leveraging Linguistic and Source-based Features for Machine Translation Evaluation. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 394-401. Baltimore, Maryland, USA.

Habash, N. and Elkholy, A. (2008). SEPIA: Surface Span Extension to Syntactic Dependency Precision-based MT Evaluation. *Proceedings of the Workshop on Metrics for Machine Translation at the meeting of the Association for Machine Translation in the Americas (AMTA-2008).* Waikiki, Hawai.

Hachey, B., Radford, W. and Curran, J. R. (2011). Graph-based named entity linking with Wikipedia in *Proceedings of the 12th International conference on Web information system engineering*, pages 213-226, Springer-Verlag. Berlin, Heidelberg.

Hamming, R.W. (1950). Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, 29 (2): 147-160.

Hamon, O. and Rajman, M. (2006). X-Score: Automatic Evaluation of Machine Translation Grammaticality. *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006)*, pp.155-160. Genoa, Italy.

He, Y., Du, J., Way, A. and van Genabith, J. (2010). The DCU Dependency-Based Metric in WMT-MetricsMATR 2010. *Proceedings of teh 5th Workshop on Statistical Machine Translation*, pp. 349-353. Uppsala, Sweeden.

Hovy, E., King, M., and Popescu-Belis, A. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, 17(1), (pp. 43–75). Springer. The Netherlands.

Hutchins J. W. and Somers H. L. (1992). *An introduction to machine translation.* London: Academic Press.

Joty, S., Guzmán, F., Márquez, L. and Nakov, P. (2014). DiscoTK: Using Discourse Structure for Machine Translation Evaluation. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 402-408. Baltimore, Maryland, USA.

Kahn, J. G., Snover, M., Ostendorf, M. (2010). Expected Dependency Pair Match: Predicting translation quality with expected syntactic structure. *Machine Translation*.

Kamp, H. (1981). A Theory of Truth and Semantic Representation. *Formal Methods in the Study of Language* (pp. 277–322). Amsterdam: Mathematisch Centrum.

Katamba, F. (1993). *Morphology*. Macmillan. London. UK.

Kendall, M. (1938). A New Measure of Rank Correlation. *Biometrika*, *30*, 81–89. Oxford Journals. UK.

Kendall, M. (1955). *Rank Correlation Methods*. Hafner Publishing Co. London, UK.

Klein, D. and Manning, Ch. D. (2003) Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430. Sapporo, Japan

Koehn, P. and Monz, Ch. (2006). Manual and automatic evaluation of machine translation between European languages. *Proceedings of NAACL 2006 Workshop on Statistical Machine Translation*, pp. 102-121. New York, USA.

Kos, K. and Bojar, O. (2009). Evaluation of Machine Translation Metrics for Czech as teh Target Language. *Prague Bulletin of Mathematical Linguistics*, 92. Prague, Czech Republic.

Kulesza, A. and Schieber, S. M. (2004). A learning approach to improving sentence level MT evaluation. *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, USA.

Landauer, T., Foltz, P. W. and Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes* 25: 259-284. Taylor and Francis. UK.

Lavie, A. and Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 228-231. Prague, Czech Republic.

Leusch, G., Ueffing, N., and Ney, H. (2006). CDER: Efficient MT Evaluation Using Block Movements. *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL),* pp. 241–248. Trento, Italy.

Leusch, G. and Ney, H. (2008). BLEUSP, INVWER, CDER: Three improved MT evaluation measures. NIST Metrics for Machine Translation Challenge. Waikiki, Honolulu, Hawai.

Leusch, G. and Ney, H. (2009). Edit distance with block movements and error rate confidence estimates. *Machine Translation*. 23:129-140. Springer. The Netherlands.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals, Soviet Physics-Doklady 10, 707-710.

Liang, P., Taskar, B. and Klein, D. (2006). Alingment by agreement. *Proceedings of the HLT-NAACL Conference*, pp. 104-111. New York, NY.

Lin, D. (1998). Dependency-based Evaluation of MINIPAR. *Proceedings of the Workshop on the Evaluation of Parsing Systems.*

Lin, C.-Y., and Och, F. J. (2004). Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. *Proceedings of the*

*42nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Columbus, Ohio, USA

Lita, L. V., Rogati, M., and Lavie, A. (2005). BLANC: Learning Evaluation Metrics for MT. *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP),* pp. 740–747.

Liu, Ch., Dahlmeier, D. and Ng, Hwee. (2010). TESLA: Translation Evaluation of Sentences with Linear-programming-based Analysis. *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pp. 354-359. Uppsala, Sweeden.

Liu, D., and Gildea, D. (2005). Syntactic Features for Evaluation of Machine Translation. *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization,* pp. 25–32.

Liu, D., and Gildea, D. (2006). Stochastic Iterative Alignment for Machine Translation Evaluation. *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL),* pp. 539–546.

Liu, D. and Gildea, D. (2007). Source-Language Features and Maximum Correlation Training for Machine Translation Evaluation. *Proceedings of the 2007 Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL),* pp. 41–48.

Lloberes, M., Castellón, I. and Padró, Ll. (2010). Spanish FreeLing Dependency Grammar. Nicoletta Calzolari et al. (ed.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 693-699. Malta.

Lo, Ch. and Wu, D. (2010). Semantic vs. Syntactic vs. N-gram Structure for Machine Translation Evalution. *Proceedings of the Fourth Workshop on Syntax and Structure in Statistical Translation*, pp. 52-60. Beijing.

Lo, Ch. and Wu, D. (2011). Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. *Proceedings of the 49th*

*Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 220–229. Portland, Oregon, USA.

Lo, Ch. and Wu, D. (2012) Unsupervised vs. Supervised weight estimation for semantic MT evaluation metrics. *The Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)*. Jeju Island, South Korea.

Lo, Ch., Tumurulu, A. K. and Wu, D. (2012). Fully Automatic Semantic MT Evaluation. *Proceedings of the 7th Workshop on Statistical Machine Translation*, pp. 243-252. Montréal, Canada.

Lo, Ch. and Wu, D. (2013). MEANT at WMT2013: A tunable, accurate yet inexpensive semantic frame based MT evaluation metric. *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pp. 422-428. Sofia, Bulgaria.

Lo, Ch., Beloucif, M., Saers, M. and Wu, D. (2014). XMEANT: Better semantic MT evaluation without reference translations. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 765-771. Maryland, USA.

MacCartney, B., Grenager, T., de Marneffe, M.C., Cer, D. and Manning C. D. (2006). Learning to recognize features of valid textual entailments. *Proceedings of NAACL*.

Macháček, M. and Bojar, O. (2011). Approximating a Deep-Syntactic Metric for MTE Evaluation and Tuning. *Proceedings of the 6th Workshop on Statistical Machine Translation*, pp. 92-98. Edinburgh, Scotland, UK.

Macháček, M. and Bojar, O. (2013). Results of the WMT14 Metrics Shared Task. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 45-51, Sofia, Bulgaria.

Macháček, M. and Bojar, O. (2014). Results of the WMT14 Metrics Shared Task. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301. Baltimore, Maryland, USA.

Manning, Ch. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, David. (2014). The Stanford CoreNLP Natural Language Processing Toolkit.

*Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Mehay, D., and Brew, C. (2007). BLEUATRE: Flattening Syntactic Dependencies for MT Evaluation. *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

Melamed, I. D., Green, R., and Turian, J. P. (2003). Precision and Recall of Machine Translation. *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL).*

Miller, G. and Fellbaum, Ch. (2007). WordNet. http://wordnet.princeton.edu/

Miller, G. A. and Fellbaum, Ch. (2007). WordNet Then and Now. *Language Resources and Evaluation*, 41:2, 209-214.

Nießen, S., Och, F. J., Leusch, G., and Ney, H. (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*.

Nivre, J., Hall J. and Nilsson, J. (2006) Maltparser: A data-driven parser-generator for dependency parsing. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pp. 2216–2219. Genoa, Italy

Olive, J. (2005). Global Autonomous Language Exploitation (GALE). DARPA/IPTO Proposer Information Pamphlet.

Owczarzak, K., Groves, D., Genabith, J. V., and Way, A. (2006). Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA),* pp. 148–155. Massachusetts, USA.

Owczarzak, K., van Genabith, J., and Way, A. (2007a). Dependency-Based Automatic Evaluation for Machine Translation. *Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation,* pp. 80–87.

Owczarzak, K., van Genabith, J., and Way, A. (2007b). Labelled Dependencies in Machine Translation Evaluation. *Proceedings of the ACL Workshop on Statistical Machine Translation,* pp. 104–111. Czech Republic.

Padó, S., Galley, M., Jurafsky, D. and Manning, Ch. D. (2009). Measuring Machine Translation Quality as Semantic Equivalence: A Metric Based on Entailment Features. *Journal of MT*, 23(2-3), pp. 181-193.

Padró, Ll. and Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012), ELRA*. Istanbul, Turkey.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). *Bleu: a method for automatic evaluation of machine translation, RC22176 (Technical Report)*. IBM T.J. Watson Research Center.

Parton, K., Tetreault, J., Madnani, N. and Chodorow, M. (2011). E-rating Machine Translation. *Proceedings of the 6th Workshop on Statistical Machine Translation*, pp. 108-115. Edinburgh, Scotland, UK.

Paul, M., Finch, A. and Sumita, E. (2007). Reducing Human Assessment of Machine Translation Quality to Binary Classifiers. *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation* (TMI).

Pearson, K. (1914, 1924, 1930). *The life, letters and labours of Francis Galton*. (3 volumes).

Popescu-Belis A., Estrella P., King M. and Underwood N. (2006). A model for context-based evaluation oflanguage processing systems and its application to machine translation evaluation. *Proceedings fourth International Conference on Language Resources and Evaluation ( LREC)*, pp.691-696. Genoa, Italy.

Popović, M. (2011). Morphemes and POS tags for n-gram based evaluation metrics. *Proceedings of the 6th Workshop on Statistical Machine Translation*, pp. 104-107. Edinburgh, Scotland, UK.

Popović, M. (2012). Class error rates for evaluation of machine translation output. *Proceedings of the 7th Workshop on Statistical Machine Translation*, pp. 71-75. Montréal, Canda.

Porter, M. (2001). The Porter Stemming Algorithm. http://www.tartarus.org/martin/PorterStemmer/index.html

Pradhan, S., Ward, W., Hacioglu, K., Martin, J. H. and Jurafsky, D. (2004). Shallow Semantic Parsing Using Support Vector Machines. *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAAACL-2004)*. Boston, MA, USA.

Quirk C. (2004) Training a sentence-level machine translation confidence measure. In: *Proceedings of the international conference on language resources and evaluation (LREC-2004)*, pp 825-828. Lisbon, Portugal.

Reeder, F., Miller, K., Doyon, J., and White, J. (2001). The Naming of Things and the Confusion of Tongues: an MT Metric. *Proceedings of the Workshop on MT Evaluation "Who did what to whom?" at Machine Translation Summit VIII*, pp. 55–59.

Rios, M., Aziz, W. and Specia, L. (2011). TINE: A metric to assess MT adequacy. *Proceedings of the 6th Workshop on Statistical Machine Translation*, pp. 116-122. Edinburgh, Scotland, UK.

Russo-Lassner, G., Lin, J., and Resnik, P. (2005). *A Paraphrase-Based Approach to Machine Translation Evaluation (LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57)* (Technical Report). University of Maryland, College Park. http://lampsrv01.umiacs.umd.edu/pubs/TechReports/LAMP\ 125/LAMP\ 125.pdf.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. *Lecture Notes in Computer Science*, Vol. 2276, pp. 1-15. Springer. The Netherlands.

Sgall, P., Hajičová, E. and Panevová, J. (1986). *The meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecth, The Netherlands.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pp. 223–231.

Snover, M., Madnani, N., Dorr, B. and Schwartz, R. (2009). Fluency, Adequacy or HTER? Exploring different human judgments with a tunable MT metric. *Proceedings of the 4th Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.

Song X. and Cohn, T. (2011). Regression and Ranking based Optimisation for Sentence Level MT Evaluation. *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp 123–129. Edinburgh, Scotland.

Song, X., Cohn, T. and Specia, L. (2013). BLEU deconstructed: Designing a better MT evaluation metric. *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*.

Spearman, C. (1904). The Proof and Measurement of Association Between Two Rings. *American Journal of Psychology*, *15*, 72–101.

Specia, L., Cancedda, N., Dymetman, M., Turchi, M. and Cristianini, N. (2009). Estimating the Sentence-Level Quality of Machine Translation Systems. *Proceedings of the 13th Annual Conference of the EAMT*, pp. 28-35. Barcelona, Spain.

Specia, L., Raj, D., Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation*. 24:39-50.

Specia, L. and Giménez, J. (2010). Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation. *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, Denver, Colorado.

Specia, L., Hajlaoui, N., Hallet, C. and Aziz, W. (2011). Prediting Machine Translation Adequacy. *Proceedings of the 13th Translation Summit*, pp. 513-520. Xiamen, China.

Stanojević, M. and Sima'an, K. (2014). Beer: Better Evaluation as Ranking. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 414-419. Baltimore, Maryland, USA.

Surdeanu, M., and Turmo, J. (2005). Semantic Role Labeling Using Complete Syntactic Analysis. *Proceedings of CoNLL Shared Task*.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated DP based Search for Statistical Translation. *Proceedings of European Conference on Speech Communication and Technology*.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work,* Amsterdam and Philadelphia: John Benjamins.

Toutanova, K., Klein, D., Manning, Dh. and Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of HLT-NAACL*, pp. 252-259.

Trask, R. L. (1992). *A Dictionary of Grammatical Terms in Linguistics*. Routledge. London.

Turian, J. P., Shen, L., and Melamed, I. D. (2003). Evaluation of Machine Translation and its Evaluation. *Proceedings of MT SUMMIT IX*.

Vilar, D., Xu, J., Fernando, D'Haro, L. F. and Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. *Proceedings of the LREC*, Genoa, Italy.

Wang, M., and Manning, Ch. (2012). SPEDE: Probabilistic Edit Distance Metrics for MT Evaluation. *Proceedings of the 7th Workshop on Statistical Machine Translation.* Montréal, Canada.

White, J. S., O'Connell, T. and O'Mara, F. (1994). The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA)*, pp. 193–205.

Wong, B. T-M. and Kit, Ch. (2008). Word choice and Word position for Automatic MT Evaluation. AMTA 2008 Workshop: MetricsMATR, Waikiki, Hawai, October.

Wong, B. T-M. and Kit, Ch. (2010). ATEC automatic evaluation of machine translation via word choice and word order. *Machine Translation* 23(2): 141-151. Springer.

Wu, X., Yu, H. and Liu, Q. (2013). DCU Participation in WMT13 Metrics Task. *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pp. 435-439. Sofia, Bulgaria.

Wu, X., Yu, H. and Liu, Q. (2014). RED: DCU-CASICT Participation in WMT14 Metrics Task. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 420-425. Baltimore, Maryland, USA.

Wu, Z. and Palmer, M. (1994). Verb Semantics and Lexical Selection. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 133-138. Las Cruces, New Mexico.

Yang, MY., Sun, SQ., Zhu, JG., Li, S., Zhao, TJ., Zhu, XN. (2011). Improvement of Machine Translation Evaluation by Simple Linguistically Motivated Features. *Journal of computer science and technology* 26(1): 57-67, January, 2011, Springer. The Netherlands.

Ye, Y., Zhou, M. and Lin, Ch. (2007). Sentence Level Machine Translation Evaluation as a Ranking Problem: one step aside from BLEU. *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 240-247. Prague, Czech Republic..

Žabokrtský, Z., Ptáček, J. and Pajas, P. (2008). TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 167–170. Columbus, OH, USA.

Zhou, L., Lin, C.-Y., and Hovy, E. (2006). Re-evaluating Machine Translation Results with Paraphrase Support. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),* pp. 77–84.

## Dictionaries

Collins English Dictionary. http://www.collinsdictionary.com/dictionary/english (last accessed: November 3rd 2014).

Macmillan English Dictionary. http://www.macmillandictionary.com/ (last accessed: November 3rd 2014).

# Appendix A. Tools and Resources

## A.1 English

The tools and resources used for English and organised per module are the following.

### A.1.1 Lexical Module

WordNet 3.0: http://wordnet.princeton.edu/

### A.1.2 Morphological Module

Standford Log-Linear PoS Tagger: http://nlp.stanford.edu/software/tagger.shtml

This parser uses the Penn Treebank English PoS Tagset:

| Tag | Description |
|---|---|
| **$** | dollar |
| **``** | opening quotation mark |
| **''** | closing quotation mark |
| **(** | opening parenthesis |
| **)** | closing parenthesis |
| **,** | comma |
| **--** | dash |
| **.** | sentence terminator |
| **:** | colon or ellipsis |
| **CC** | conjunction, coordinating |
| **CD** | numeral, cardinal |
| **DT** | determiner |
| **EX** | existential there |
| **FW** | foreign word |
| **IN** | preposition or conjunction, subordinating |
| **JJ** | adjective or numeral, ordinal |

| | |
|---|---|
| **JJR** | adjective, comparative |
| **JJS** | adjective, superlative |
| **LS** | list item marker |
| **MD** | modal auxiliary |
| **NN** | noun, common, singular or mass |
| **NNP** | noun, proper, singular |
| **NNPS** | noun, proper, plural |
| **NNS** | noun, common, plural |
| **PDT** | pre-determiner |
| **POS** | genitive marker |
| **PRP** | pronoun, personal |
| **PRP$** | pronoun, possessive |
| **RB** | adverb |
| **RBR** | adverb, comparative |
| **RBS** | adverb, superlative |
| **RP** | particle |
| **SYM** | symbol |
| **TO** | "to" as preposition or infinitive marker |
| **UH** | interjection |
| **VB** | verb, base form |
| **VBD** | verb, past tense |
| **VBG** | verb, present participle or gerund |
| **VBN** | verb, past participle |
| **VBP** | verb, present tense, not 3rd person singular |
| **VBZ** | verb, present tense, 3rd person singular |
| **WDT** | Wh-determiner |
| **WP** | Wh-pronoun |
| **WP$** | Wh-pronoun, possessive |

| | |
|---|---|
| **WRB** | Wh-adverb |

**Table A.1** Penn Treebank Tagset

### A.1.3 Dependency Module

The PCFG parser for English (Lein and Manning 2003; de Marneffe et al. 2006) contained in the Standford CoreNLP suite: http://nlp.stanford.edu/software/lex-parser.shtml#Download

The dependency labels used by this parser are the following[57]:

| Label | Description |
|---|---|
| **abbrev** | abbreviation modifier |
| **acomp** | adjectival complement |
| **advcl** | adverbial clause modifier |
| **advmod** | adverbial modifier |
| **agent** | agent |
| **amod** | adjectival modifier |
| **appos** | appositional modifier |
| **attr** | attributive |
| **aux** | auxiliary |
| **auxpass** | passive auxiliary |
| **cc** | coordination |
| **ccomplm** | clausal complement with internal subject |
| **comp** | complement |
| **conj** | conjunct |
| **cop** | copula |
| **csubj** | clausal subject |
| **csubjpass** | clausal passive subject |
| **dep** | dependent |
| **det** | determiner |
| **discourse** | discourse element (e.g. fillers, interjections) |

[57] For further description of the dependency types see Marneffe and Manning (2008).

| | |
|---|---|
| **dobj** | direct object |
| **expl** | expletive (expletive there) |
| **infmod** | infinitival modifier |
| **iobj** | indirect object |
| **mark** | marker (word introducing an advcl) |
| **measure** | measure |
| **mod** | modifier |
| **mwe** | multi-word expression |
| **neg** | negation modifier |
| **nn** | noun compound modifier |
| **npadvmod** | noun phrase as adverbial modifier |
| **nsubj** | nominal subject |
| **nsubjpass** | passive nominal subject |
| **num** | numeric modifier |
| **number** | element of compound number |
| **parataxis** | parataxis |
| **partmod** | participial modifier |
| **pcomp** | prepositional complement |
| **pobj** | object of a preposition |
| **poss** | possession modifier |
| **possessive** | possessive modifier ('s) |
| **preconj** | preconjunct |
| **predet** | predeterminer |
| **prep** | prepositional modifier (also pmod, sometimes) |
| **prepc** | prepositional clausal modifier |
| **prt** | phrasal verb particle |
| **punct** | punctuation |
| **purpcl** | purpose clause modifier |
| **quantmod** | 1uantifier phrase modifier |
| **rcmod** | relative clause modifier |
| **ref** | referent |
| **rel** | relative (word introducing a rcmod) |
| **root** | root |

| | |
|---|---|
| **tmod** | temporal modifier |
| **xcomp** | clausal complement with external subject |
| **xsubj** | controlling subject |

**Table A.2** Dependency labels and their description

### A.1.4 Semantic Module

The tools used in the Semantic Module are:

- Supersense Tagger to recognize NEs: http://sourceforge.net/projects/supersensetag/
- NE Linking: a graph-based NEL tool inspired by Hachey et al. (2011).
- The Stanford Temporal Tagger to identify and normalize Time Expressions: http://nlp.stanford.edu/software/sutime.shtml
- Sentiment analysis tool described in Atserias et al. (2012).

### A.1.5 Language Model Module

A News Language Model: http://www.quest.dcs.shef.ac.uk/quest_files/de-en/news.3gram.en.lm

## A.2 Spanish

The tools and resources used for Spanish and organised per module are the following.

### A.2.1 Lexical Module

- WordNet 3.0: http://grial.uab.es/synset/synset2.php
- Freeling Lemmatizer: http://nlp.lsi.upc.edu/freeling/index.php?option=com_content&task=view&id=13&Itemid=42

### A.2.2 Morphological Module

PoS Tagger Module in Freeling: http://nlp.lsi.upc.edu/freeling/doc/userman/html/node48.html

The PoS tags used in Freeling are EAGLES PoS tags, which are described in http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

### A.2.3 Dependency Module

Txala parser in Freeling: http://nlp.lsi.upc.edu/freeling/doc/userman/html/node59.html

The dependency labels provided by the Freeling dependency parser are the following:

| Label | Description |
|---|---|
| **adj-mod** | adjectives modifying their head |
| **ador** | sentence adjunct |
| **agent** | agent - passive |
| **att** | predicate whose head is a copulative verb |
| **aux** | auxiliary verbs |
| **cc** | adjunct |
| **co-adj** | adjective coordination |
| **co-adv** | adverb coordination |
| **co-ger** | gerund coordination |
| **co-inf** | infinitive coordination |
| **co-n** | noun coordination |
| **co-part** | participle coordination |
| **co-sp** | prepositional phrase coordination |
| **co-subord** | clause coordination |
| **co-v** | verb or sentence coordination |
| **dconj** | verb + conjunction - periphrasis |
| **dep-adv** | adverbs - verbless sentences |
| **dep-ger** | gerund clauses - verbless sentences |
| **dep-inf** | infinitive clauses - verbless sentences |
| **dep-noun** | nouns - verbless sentences |
| **dep-part** | participle clauses - verbless sentences |
| **dep-prep** | prepostitional phrases - verbless sentences |
| **dep-subord** | finite clauses - verbless sentences |
| **dep** | clitics |
| **dobj** | direct object |
| **dprep** | verb + preposition - periphrasis |
| **dverb** | verb + verb - periphrasis |
| **es** | passive, impersonal, pronominal morpheme, reflexive pronouns |

| | |
|---|---|
| **espec** | nominal and verbal determiners |
| **iobj** | indirect object |
| **pred** | second predicate whose head is other than copulative verbs |
| **obj-prep** | objects whose head is prepositional |
| **sn-mod** | noun phrases modifying their head |
| **sp-mod** | preposition phrases modifying their head |
| **sp-obj** | prepositional object |
| **subj-pac** | patient - passive |
| **subj** | subject |
| **subord-mod** | relative clauses modifying their head |
| **term** | punctuation |
| **top** | sentence head (highest head) |
| **vsubord** | conjunction + verb - subordinate clauses |

**Table A.3** Dependency labels and their description

# Appendix B. Acronyms and Abbreviations

| | |
|---|---|
| ACL | Association for Computational Linguistics |
| Adj | Adjunct |
| Adj-Time | Adjunct of Time |
| ALPAC | Automatic Language Processing Advisory Committee |
| ARPA | Advanced Research Projects Agency |
| Cs | Subject Complement |
| cz | Czech |
| de | German |
| DRS | Discourse Representation Structures |
| DRT | Discourse Representation Theory |
| en | English |
| es | Spanish |
| FEMTI | Framework for Machine Translation Evaluation in ISLE |
| fr | French |
| hi | Hindi |
| HYP | Hypothesis |
| ISLE | International Standards for Language Engineering |
| KNOW2 | Language understanding technologies for multilingual domain-oriented information access |
| LFG | Lexical Functional Grammar |
| LM | Language Model |

| | |
|---|---|
| MetricsMATR | Metrics for Machine Translation |
| MFS | Most Frequent Word Sense |
| MT | Machine Translation |
| NE | Named Entity |
| NEL | Named Entity Linking |
| NER | Named Entity Recognition |
| NIST | National Institute of Standards and Technology |
| NLP | Natural Language Processing |
| Non-fin Cl | Non-finite Clause |
| NP | Noun Phrase |
| NP-Gen | Noun Phrase Genitive |
| Obl | Oblique |
| Od | Direct Object |
| OpenMT | Open Machine Translation |
| OSV | Object-Subject-Verb |
| OVS | Object-Verb-Subject |
| PCFG | Probabilistic Context-free Grammar |
| PoS | Part-of-Speech |
| Post-mod | Post-modifier |
| PP | Prepositional Phrase |
| QE | Quality Estimation |
| REF | Reference |

| | |
|---|---|
| RTE | Recognizing Textual Entailment |
| ru | Russian |
| SOV | Subject-Object-Verb |
| SR | Semantic Role |
| SRL | Semantic Role Labelling |
| Subj | Subject |
| SVClCompl | Subject-Verb-Clause Complement |
| SVM | Super Vector Machine |
| SVO | Subject-Verb-Object |
| SVObl | Subject-Verb-Oblique |
| TC-STAR | Technology and Corpora for Speech to Speech Translation |
| TE | Textual Entailment |
| tf-idf | frequency-inverse document frequency |
| TIMEX | Time Expressions |
| ULC | Uniformly-averaged Linear Combination |
| VOS | Verb-Object-Subject |
| VP | Verb Phrase |
| VSO | Verb-Subject-Object |
| WMT | Workshop on Statistical Machine Translation |
| XCompl | X-Complement |
| XML | Extensible Markup Language |

# Appendix C. Summary of the Metrics Using Linguistic Information

| MT Metrics | Linguistic Features | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Stem. | Syn. | Para. | Funct. & Cont. Words | Spel. | Morph. Info | Lem. | PoS | Const. | Dep. | LM | NE | SR | DR | TE | Sem PoS | Thes | Ont |
| METEOR | X | X | X | X | | | | | | | | | | | | | | |
| M-TER | X | X | | | | | X | X | | | | | | | | | | |
| M-BLEU | X | X | | | | | X | X | | | | | | | | | | |
| ATEC | X | X | | | | | X | X | | | | | | | | | | |
| TERp | X | X | X | | | | X | X | | | | | | | | | | |
| INVWER | | | | | X | X | | | | | | | | | | | | |
| CDER | | | | | X | X | | | | | | | | | | | | |
| AMBER | | | | | | X | | | | | | | | | | | | |
| SPEDE | | X | X | | | | | X | | | | | | | | | | |
| INFER | | | | | | X | | | | | | | | | | | | |
| SMT | | | | | | | | | X | | | | | | | | | |
| HWCM | | | | | | | | | | X | | | | | | | | |
| X-score | | | | | | | | X | X | | | | | | | | | |
| BLEUÂTRE | | | | | | | | | | X | | | | | | | | |
| Owczarzak | | X | | | | | | | | X | | | | | | | | |
| SEPIA | | | | | | | | | | X | | | | | | | | |
| SP | | | | | | | | X | | | | | | | | | | |
| CP | | | | | | | | X | X | | | | | | | | | |
| DP | | | | | | | | | | X | | | | | | | | |
| EDPM | | | | | | | | | | X | | | | | | | | |
| DCU-Dep | X | X | X | | | | | | | X | | | | | | | | |
| TESLA-M | | | | X | | | | X | | | | | | | | | | |
| TESLA-B | | | X | X | | | | X | | | | | | | | | | |
| TESLA-F | | | X | X | | | | X | | | X | | | | | | | |

| **MT Metrics** | **Linguistic Features** | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Stem. | Syn. | Para. | Funct. & Cont. Words | Spel. | Morph. Info | Lem. | PoS | Const. | Dep. | LM | NE | SR | DR | TE | Sem PoS | Thes | Ont |
| POSF | | | | | | | | X | | | | | | | | | | |
| MPF | | | | | | X | | X | | | | | | | | | | |
| WMPF | | | | | | X | | X | | | | | | | | | | |
| TerrorCat | | | | | | | | X | | | | | | | | | | |
| DepRef | X | X | | | | | | | | X | | | | | | | | |
| RED | X | X | X | X | | | | | | X | | | | | | | | |
| NEE | | | | | | | | | | | | X | | | | | | |
| NE | | | | | | | | | | | | X | | | | | | |
| SR | | | | | | | | | | | | | X | | | | | |
| DR | | | | | | | | | | | | | | X | | | | |
| RTE | | | | | | | | | | | | | | | X | | | |
| SemPoS | | | | | | | | | | | | | | | | X | | |
| SAGAN-STS | | | | | | | | | | | | | | | X | | | |
| MEANT | | | | | | | | | | | | | X | | | | | |
| Akiba et al. 2001 | X | | | | | | | X | | | | | | | | | X | |
| Paul et al. 2007 | X | X | X | X | | | | | | | | | | | | | | |
| Albrecht & Hwa 2007 | X | X | X | X | | | | | X | X | | | | | | | | |
| Ye et al. 2007 | | | | | | | | | | X | X | | | | | | | |

| **MT Metrics** | **Linguistic Features** | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Stem. | Syn. | Para. | Funct. & Cont. Words | Spel. | Morph. Info | Lem. | PoS | Const. | Dep. | LM | NE | SR | DR | TE | Sem PoS | Thes | Ont |
| CD6P4ER | | | | | X | X | | | | | | | | | | | | |
| Yang et al. 2011 | | | | X | | | | | X | | | | | | | | | |
| TINE | X | | | | | | | | | | | | X | | | | | X |
| Layered | | | | | | | | | | X | | | | X | | | | |
| DiscoTK | | X | X | | | | | X | X | X | | X | X | X | | | | |
| MAXSIM | | X | | | | | X | X | | X | | | | | | | | |
| Giménez & Márquez 2010 | X | X | | | | | | X | X | X | | | X | X | | | | |
| González et al. 2014 | X | X | X | | | | X | X | X | X | X | X | X | X | | | | |