The Role of Human Reference Translation in Machine Translation Evaluation

Marina Fomicheva

TESI DOCTORAL UPF / ANY 2017

DIRECTOR DE LA TESI
Prof. Núria Bel
Prof. Iria da Cunha
Departament de Traducció i Ciències del Llenguatge

To Anton

Acknowledgments

While working on this thesis, I received help and encouragement from many people. First and foremost, I would like to thank my supervisors, Núria Bel and Iria da Cunha. They supported me with great involvement and care while allowing me to pursue my ideas.

A special thank you goes to Lucia Specia for making me feel welcome in Sheffield, for her insightful ideas, constructive comments and fruitful collaboration. I would also like to thank the members of Sheffield MT group for being such great colleagues during my research stay.

I am grateful to the members of the IULA and my PhD colleagues for the support and inspiration that they have given me during these years.

I would also like to thank my friends, here in Barcelona, in Saint-Petersburg and in Madrid for cheering me up. A special thank you goes to Jenny, for making me feel at home during my first year in Barcelona and to Polina for all the fun and fencing.

I am deeply thankful to my parents, Vladimir and Victoria, for their unconditional love and support.

Finally, my most special thank you goes to Anton for his genuine interest and help, for inspiring discussions and for passing me his can-do attitude. Without you it would not have been possible!

This thesis was developed in the framework of the project TUNER TIN2015-65308-C5-1-R (MINECO/FEDER, UE), with the financing of the FI-DGR2014 Scholarship of the Secretary of Universities and Investigation (Secretaria de Universidades e Investigacion, SUR) of the Department of Economy and Knowledge (Departamento de Economıa y Conocimiento, DEC) of the Generalitat de Catalunya and the IULA-UPF Scholarship of the Institute for Applied Linguistics (Institut Universitari de Linguıstica Aplicada, IULA).

Abstract

Both manual and automatic methods for Machine Translation (MT) evaluation heavily rely on professional human translation. In manual evaluation, human translation is often used instead of the source text in order to avoid the need for bilingual speakers, whereas the majority of automatic evaluation techniques measure string similarity between MT output and a human translation (commonly referred to as candidate and reference translations), assuming that the closer they are, the higher the MT quality.

In spite of the crucial role of human reference translation in the assessment of MT quality, its fundamental characteristics have been largely disregarded. An inherent property of professional translation is the adaptation of the original text to the expectations of the target audience. As a consequence, human translation can be rather different from the original text, which, as will be shown throughout this work, has a strong impact on the results of MT evaluation.

The first goal of our research was to assess the effects of using human translation as a benchmark for MT evaluation. To achieve this goal, we started with a theoretical discussion of the relation between original and translated texts. We identified the presence of optional translation shifts as one of the fundamental characteristics of human translation. We analyzed the impact of translation shifts on automatic and manual MT evaluation showing that in both cases quality assessment is strongly biased by the reference provided.

The second goal of our work was to improve the accuracy of automatic evaluation in terms of the correlation with human judgments. Given the limitations of reference-based evaluation discussed in the first part of the work, instead of considering different aspects of similarity we focused on the differences between MT output and reference translation searching for criteria that would allow distinguishing between acceptable linguistic variation and deviations induced by MT errors. In the first place, we explored the use of local syntactic context for validating the matches between candidate and reference words. In the second place, to compensate for the lack of information regarding the MT segments for which no counterpart in the reference translation was found, we enhanced reference-based evaluation with fluency-oriented features. We implemented our approach as a family of automatic evaluation metrics that showed highly competitive performance in a series of well known MT evaluation campaigns.

Resumen

Tanto los métodos manuales como los automáticos para la evaluación de la Traducción Automática (TA) dependen en gran medida de la traducción humana profesional. En la evaluación manual, la traducción humana se utiliza a menudo en lugar del texto original para evitar la necesidad de hablantes bilingües, mientras que la mayoría de las técnicas de evaluación automática miden la similitud entre la TA y una traducción humana (comúnmente llamadas traducción candidato y traducción de referencia), asumiendo que cuanto más cerca están, mayor es la calidad de la TA.

A pesar del papel fundamental que juega la traducción de referencia en la evaluación de la calidad de la TA, sus características han sido en gran parte ignoradas. Una propiedad inherente de la traducción profesional es la adaptación del texto original a las expectativas del lector. Como consecuencia, la traducción humana puede ser bastante diferente del texto original, lo cual, como se demostrará a lo largo de este trabajo, tiene un fuerte impacto en los resultados de la evaluación de la TA.

El primer objetivo de nuestra investigación fue evaluar los efectos del uso de la traducción humana como punto de referencia para la evaluación de la TA. Para lograr este objetivo, comenzamos con una discusión teórica sobre la relación entre textos originales y traducidos. Se identificó la presencia de cambios de traducción opcionales como una de las características fundamentales de la traducción humana. Se analizó el impacto de estos cambios en la evaluación automática y manual de la TA demostrándose en ambos casos que la evaluación está fuertemente sesgada por la referencia proporcionada.

El segundo objetivo de nuestro trabajo fue mejorar la precisión de la evaluación automática medida en términos de correlación con los juicios humanos. Dadas las limitaciones de la evaluación basada en la referencia discutidas en la primera parte del trabajo, en lugar de enfocarnos en la similitud, nos concentramos en el impacto de las diferencias entre la TA y la traducción de referencia buscando criterios que permitiesen distinguir entre variación lingüística aceptable y desviaciones inducidas por los errores de TA. En primer lugar, exploramos el uso del contexto sintáctico local para validar las coincidencias entre palabras candidato y de referencia. En segundo lugar, para compensar la falta de información sobre los segmentos de la TA para los cuales no se encontró ninguna relación con la traducción de referencia, introdujimos características orientadas a la fluidez de la TA en la evaluación basada en la referencia. Implementamos nuestro enfoque como una familia de métricas de evaluación automática que mostraron un rendimiento altamente competitivo en una serie de conocidas campañas de evaluación de la TA.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Automatic evaluation of Machine Translation (MT) is a very challenging task that aims to emulate human assessment of MT quality. In spite of the important advances, existing automatic evaluation metrics still fail to accurately approximate the results of manual evaluation. An anecdotal remark by Yorik Wilks that "more has been written about MT evaluation than about MT itself" (King et al. (2003, p.224)) may be an exaggeration but it reflects the fact that, automatic evaluation has become an active research field in its own right.

Both manual and automatic methods heavily rely on professional human translation using it as a benchmark for evaluation. In manual evaluation, MT output is often compared to a human translation instead of the original text in order to avoid the need for bilingual speakers, whereas many of the well known automatic evaluation techniques rely on string similarity between MT output and a human translation, assuming that the closer they are, the higher the MT quality.

In spite of the crucial role played by human translation in MT evaluation, its fundamental characteristics have been largely disregarded. An inherent property of professional translation is the adaptation of the original text to the expectations of target language audience that depend on text type and genre, cultural conventions, situational context, etc. As a consequence, human translation can be rather different from the original text, which in our view, has a strong impact on the results of MT evaluation.

In this thesis we study the implications of using human translation as a benchmark for assessing MT quality. Starting from a discussion of the challenges of MT evaluation in the light of translation studies, we conduct an empirical investigation of the impact of the characteristics of human translation on the results of MT evaluation and propose various practical solutions to the related problems.

1.1 Machine Translation Evaluation

The definition of translation quality in itself is a controversial issue. As pointed out by House (2009), "translation quality presupposes a theory of translation" (House (2009, p.222)). Different ways of understanding quality in the field of translation studies include response-oriented (Nida, 1964/2000; Gutt, 2014), text-based (Reiss, 2014) and functional-pragmatic approaches (House, 2001), among others. Different views on translation itself lead to different concepts of translation quality and different ways of assessing it.[1]

A common ground in the various definitions of translation quality is that "by its very nature, translation is simultaneously bound to the source text and the presuppositions and conditions governing its reception in the target linguistic and cultural systems" (House (2009, p.224)). Translation quality is, therefore, determined by the tension between two prototypical expectations: that of fidelity to the original text and that of naturalness of expression in the target language.

Modern theories of translation emphasize its communicative aspects and the role of translated text in the context of receiving target culture (Toury, 2012). Besides the source text, translation choices are conditioned by a variety of extra-linguistic factors, such as the target audience, cultural differences, background knowledge, language-specific genre and text type conventions, etc. Translation criticism, i.e. human quality assessment, typically assesses the adequacy of the choices made by the translator in the light of these factors.

Naturally, the perspectives on the quality of professional human translation and on MT quality are quite different. Translation criticism with all its complexity is hardly suitable for MT evaluation. In the field of MT, there are two basic approaches to translation quality, sometimes referred to as extrinsic (or task-oriented) and intrinsic quality measures (Dorr et al., 2011). On the one hand, the fact that MT can hardly substitute a human translation and is still used in a limited number of scenarios gives rise to the task-oriented evaluation, where the MT output is assessed in terms of different user-focused aspects, such as reliability of the software, efficiency and speed, usability or suitability for a particular task (Hovy et al., 2002). For example, if the task in mind is post-editing, one can use post-editing time as a measure of quality (Specia, 2011). If MT is intended for gisting purposes, reading comprehension tests can be used for evaluation (Church and Hovy, 1993).

On the other hand, in spite of its limitations, MT still is a rendering of the contents of the source text in the target language, and therefore, in principle can be evaluated with the measures applied to human translation quality, such as fidelity to the original and naturalness of expression in the target language. This type of evaluation typically

[1] The discussion of existing models of translation quality is beyond the scope of this thesis. The reader is referred to House (2009) for an overview of existing approaches.

takes the form of subjective human judgments on a multi-point scale. The most common approach developed back in the nineties by Advanced Research Project Agency (ARPA) (White et al., 1994) and still widely used in the field follows the methodology originally designed for the evaluation of human translation but with substantial simplifications. Initially, the method included the assessment of lexical, grammatical, semantic and stylistic aspects by a panel of professional translators. However, the errors in MT are quite different from the ones that can be found in professional human translation (Farrús et al., 2010). Furthermore, this approach was difficult to deploy logistically, as it is very hard to get a sufficient number of language-pair-specific translation experts to commit their time to such an effort. For these reasons, quality panel approach was abandoned in favor of more intuitive judgments that can be elicited from non-experts. MT outputs are typically judged on a multi-point scale in terms of adequacy (fidelity of the MT to the original text) and fluency (conformance to the norms of the target language). The judgments collected in this way can be highly subjective, as individual annotators have different backgrounds and varying perceptions and expectations regarding translation quality. In addition, to further simplify the evaluation task and avoid the need for bilingual speakers, a professional human translation is typically used instead of the source text as a benchmark for comparison.

Traditionally, manual evaluation was the only means of providing the necessary feedback for MT development. Manual evaluation, however, is slow, expensive, and subjective, whereas the development of modern MT systems requires frequent and consistent evaluation of large amounts of data. This motivated the elaboration of methods for automatic MT evaluation that aim to emulate human judgments. Although automatic evaluation is an imperfect substitute for manual quality assessment, it is now widely used in both translation industry and research, to the point that automatic evaluation metrics guide MT development as "MT researches design their systems based on the rise and fall of automatic evaluation scores" (Koehn (2009, p.222)).

Automatic MT evaluation is based on the idea that the closer the MT output is to a human reference translation, the higher its quality. Thus, the task is approached by measuring some kind of similarity between the MT output and one or various human translations of the same original text (commonly referred to as candidate and reference translations, respectively). Typically, evaluation metrics follow a simple strategy of counting the number of matching words and word strings in the MT and reference translations. For example, the most widely used reference-based evaluation metric, BLEU (Papineni et al., 2002), measures the number of word n-grams in the MT output that are also present in one or various references. Following the initial enthusiasm, n-gram-matching strategy has been severely criticized. A seminal paper by Callison-Burch and Osborne (2006) that motivated the research in automatic MT evaluation in the following years, shows that an improvement in BLEU scores is neither necessary nor sufficient for achieving an improvement in the actual MT quality. They prove that,

on the one hand, n-gram matching strategy allows for unacceptable candidate-reference variations assigning the same score to translations very different in their quality, and on the other hand, high quality translations that contain different words but express the same meaning are underestimated by the metric.

After the limitations of n-gram-based approaches had been realized, the development of evaluation metrics largely focused on how to make a more intelligent comparison between candidate and reference translations. For example, an important body of work addresses acceptable linguistic variation, so that words that have the same meaning are not considered a mismatch (for example, using synonym and paraphrase resources). Another strategy consists in measuring similarity at different levels of linguistic representation (lexical, morphological, syntactic, semantic and even discourse). Thus, we can find metrics that compute structural similarity using syntactic representations (Liu and Gildea, 2005), metrics exploiting the similarity over named entities and predicate-argument structures (Lo et al., 2012) and discourse-based metrics (Guzmán et al., 2014). Finally, further improvements have been recently achieved by combining these partial measurements using different strategies including machine learning techniques (Guzmán et al. (2014); Stanojevic and Sima'an (2014); Yu et al. (2015), *inter alia*). See Chapter 2 for a description of the current approaches to automatic MT evaluation.

Since automatic evaluation aims to emulate manual quality assessment, the performance of evaluation metrics is assessed in terms of the correlation between their scores and human judgments. Depending on the purpose of evaluation, different levels of detail may be required. MT can be evaluated at system level, i.e. providing a single measurement for a set of sentences generated by an MT system, or at sentence level, where a score for each sentence needs to be provided. System-level evaluation is useful for comparing the performance of different MT systems and allows identifying the advantages and limitations of MT strategies. Sentence-level evaluation provides fine-grained judgments of translation quality. It is necessary for exploratory error analysis and is required by many state-of-the-art algorithms for discriminative training of statistical MT systems (Liang et al. (2006), Chiang et al. (2008), Hopkins and May (2011), *inter alia*). It must be noted that simple n-gram-based methods can attain a relatively high correlation at system level. The number of matching words and phrases in system and human translations may be a fair predictor of quality if averaged over many sentences. While system-level evaluation is largely considered a solved problem, sentence-level assessment still leaves large room for improvement.

In our view, the accuracy of reference-based automatic evaluation is hindered by two related issues. In the first place, translation is an open task and therefore, the differences between an MT output and a particular human translation are not necessarily indicative of MT errors. In principle, the contents of the source text can be expressed in many different ways, the optimal choices depending on multiple linguistic and extra-linguistic

factors. Guided by these factors human translation can be rather different from the source text, which as we will argue throughout this thesis, puts in question the reliability of reference-based evaluation. In the second place, even when the differences between MT output and the reference indeed indicate the presence of errors in the former, their impact on human perception of translation quality varies greatly. As shown by Kirchhoff et al. (2014), even if MT errors are correctly identified the number of errors in a sentence is not a reliable indicator of MT quality as perceived by human judges. Thus, one of the major difficulties in reference-based evaluation stems from the fact that there is no straightforward way to factorize sentence-level scores into the counts of candidate-reference differences or similarities, since the validity of the matches, as well as the impact of the differences, varies depending on the underlying MT error type, position of the words involved, their function in the sentences, etc.

Source:	Ламб сказал диспетчеру, что полиции необходимо послать людей к нему домой.
Ref*:	Lamb told dispatcher that police needs to send people to him home.
Ref:	Lamb told the dispatcher that police needed to send officers over to his home.
MT1:	Lamb told the dispatcher that the police needs to send people to him home.
MT2:	Lamb told the dispatcher that the police need to send people to him.
MT3:	Lamb said the Dispatcher that police need to send people to his home.
MT4:	Lamb said to dispatcher that it is necessary to send to the police people to it home.

Table 1.1: Example of candidate-reference differences (WMT16 dataset, Russian-English translation, sentence 688)

As an illustration of some of the challenges of reference-based evaluation addressed in this work, consider the example in Table 1.1.[2] First of all, intuitively, if asked to evaluate the quality of the MT outputs based on the reference translation, we would note that not all the differences are equally important. For instance, the omission of the word "home" in MT2 appears as a less serious error than the displacement of the subject group "the police" in MT4. We suggest that the effect of candidate-reference differences depends not only on the underlying type of MT error, but also on the function and position of the words involved (e.g. the subject of a sentence vs. a determiner),

[2]Unless indicated otherwise the examples in this thesis are extracted from real datasets. Here and in the rest of the examples "Source" indicates the original sentence, "Ref" indicates human reference translation and "MT" indicates the corresponding MT outputs. In "Ref*" we provide a close translation into English which preserves the linguistic properties of the original sentence as closely as possible introducing only the changes that are strictly necessary to avoid violating the norms of the target language.

the context in which they occur (e.g. a word that is completely out of context vs. a word that is different from the original meaning but still makes sense in the context of the sentence), the degree of divergence from the original meaning (a related word vs. a completely different word), the linguistic aspect involved (correct words but wrong order vs. correct order but wrong words), etc. Second, if we compare the MT outputs to a close translation of the original sentence (Ref* in Table 1.1), we will see that the difference between the words "people" and "officers" is, in fact, related to an explicitation shift (i.e. the use of a word with a more specific meaning) in the reference translation and not to an MT error.

Note that if MT outputs are evaluated based on the number of matching candidate and reference words, all the above distinctions will have very similar impact of the resulting evaluation scores.

1.2 Human Translation vs. Machine Translation

Vast amount of work on automatic MT evaluation has focused on devising different ways of measuring the similarity between MT output and human translation. In spite of the crucial role that human translation plays in MT evaluation, its characteristics have been largely disregarded in the discussion of evaluation methods. These questions are in the center of attention in the field of translation studies, an academic discipline dealing with the systematic study of the theory, description and application of translation (Holmes, 1988). As pointed out by Hardmeier (2014),

> While it might seem that there should be strong connections between the two research areas [MT and translation studies], even a superficial look at the relevant literature quickly reveals that the two fields are preoccupied with completely different problems. In translation studies, much work has been devoted to defining and exploring the nature of translation. [...] Much of the SMT research literature is fairly technically-minded and is concerned with finding more effective ways of applying existing statistical methods and techniques to the MT task without spending too much thought on the effects of using these methods on perceived translation quality. (Hardmeier (2014, p.13–14))

In translation the content of the original text needs to be re-expressed using linguistic means of the target language that often conventionalizes conceptual categories in different ways. This "non-isomorphism" of source and target linguistic constructions leads to the inevitable gains and losses in translation (Szymańska, 2011) giving rise to a multiplicity of translation choices regarding the ways in which the equivalence between the original and translated texts can be achieved. Most often than not professional translators strive for the so called dynamic equivalence (Nida, 1964/2000), i.e. the equivalence

of effect that the original and translated texts are expected to have on the respective source and target audience. This involves changing the original far beyond strict necessity in order to adapt it to the linguistic regularities and cultural conventions of the target language. In fact, as we will see in Chapter 3, translation can differ from the original in any linguistic aspect (lexical, syntactic or semantic) and still be considered perfectly acceptable under given conditions.

The changes that professional translators introduce to the original text are referred to as translation shifts and have long been the focus of attention in the field of translation studies (Catford, 1965; van Leuven-Zwart, 1989; Cyrus, 2009). While some translation shifts are mandatory as they are dictated by typological differences between the languages involved, others are optional in the sense that their absence does not lead to agrammaticality and does not distort the meaning of the original. For example, if the potential reader of the translated text lacks the background knowledge to infer the information left implicit in the original, the translator is expected to express this information explicitly. As a matter of fact, some translation scholars claim that explicitation, along with simplification and normalization occur in human translation as a result of intrinsic characteristics of cognitive processes involved in translation, as well as the role of translator as mediator between source and target cultures (Baker, 1993, 1996) (see Section 3.1.2). Thus, human translation is sub-determined by the original text, i.e. translation choices cannot be exhaustively explained based on the original text and typological differences between source and target languages.

MT holds a much closer relationship to the original text. In broad terms, MT strategies can be divided in rule-based and data-driven approaches. Rule-based systems operate based on a set of manually created linguistic rules and bilingual dictionaries. Data-driven systems, with phrase-based statistical MT being the most well known approach, infer translation knowledge from large monolingual and parallel corpora.[3] Statistical MT proceeds based on translation probability statistics collected from parallel corpora (translation model) and target language n-gram frequency statistics (language model), translating one small word sequence at a time and making harsh independence assumptions. Thus MT is largely limited by the information explicitly expressed in the original sentences with no access to situational context, background knowledge or common sense reasoning involved in human translation. As we will see in the following Chapters, translation choices in the reference that are conditioned by extra-linguistic factors introduce noise in MT evaluation. While it can be argued that MT ultimately aims at delivering the same results as professional translation, some of the differences are not a consequence of MT errors but rather an artifact of the characteristics of human translation inaccessible to the current MT engines. In our view, a proper understanding of this

[3] In this work we mostly refer to data-driven systems as they constitute a vast majority in the available datasets that we use for our experiments. A detailed description of MT strategies is beyond the scope of our work. For a detailed overview of rule-based MT we refer the reader to (Hutchins and Somers, 1992), whereas a thorough presentation of statistical MT can be found in Koehn (2009).

issue will help correctly establishing priorities in MT development.

Some concerns have already been raised regarding the use of general-purpose human translation as a gold standard for the evaluation of MT quality (Ahrenberg, 2005; Friedman et al., 2008). For example, in regards to the creation of gold standard translations for the well known DARPA GALE program, "Creating human translations for the express purpose of MT evaluation requires different standards and priorities than human translation created for general use" (Friedman et al. (2008, p.1)). In our opinion, however, this issue has not received due attention in research on MT evaluation.

1.3 Goals

Before introducing the goals of our research, some general observations regarding the scope of our work are in place. First, of the two general approaches to MT evaluation mentioned above, extrinsic and intrinsic (Dorr et al., 2011), we concentrate on the latter. Second, the main focus of this work is on automatic MT evaluation at sentence level, since as mentioned in the previous Section, system-level evaluation can be largely considered a solved problem. Third, following the common settings for the development of automatic evaluation metrics, we treat manual evaluation as gold standard and assess the performance of our approach to evaluation in terms of the correlation with human judgments. However, as will be seen below, we dedicate a part of this thesis to the investigation of the problems of manual evaluation. The consistency and reliability of manual evaluation is critical for a meaningful comparison between different automatic evaluation strategies and its analysis and improvement must be part of the research in automatic MT evaluation.

Our first goal is to assess the effects of using human translation as a benchmark for MT evaluation. Both manual and automatic MT evaluation methods rely on human translation for assessing MT quality. We discuss relevant notions from translation studies concerning the relation between translated and original texts to provide a better understanding of the characteristics of human translation that affect reference-based evaluation. We have defined the following specific objectives, focusing on the role of reference translation in automatic and manual evaluation:

- **To analyze the effects of translation shifts on the results of automatic MT evaluation.** As a proof-of-concept experiment, we compare the results of automatic evaluation in the presence / absence of translation shifts in the reference. The creation of close translation variants (i.e. translation variants that do not contain optional changes with respect to the original sentences) is addressed as a rule-based paraphrase generation task. This part of our work is focused on the syntactic aspect of acceptable linguistic variation in the English-Spanish translation direction.

- **To demonstrate the influence of reference bias on manual MT evaluation.** By contrast to automatic evaluation metrics, human annotators are assumed to recognize acceptable variation between candidate and reference translations and assign a high score to a high-quality MT even if it is different from the available reference. We test this assumption by conducting an experiment where the same set of MT outputs is manually assessed using different reference translations.

The second goal of our work is to improve the accuracy of automatic evaluation in terms of correlation with human judgments. We devise various strategies to better assess the impact of individual candidate-reference differences on the perception of sentence-level MT quality by human judges. We have established the following specific objectives:

- **To develop a new evaluation metric capable of distinguishing between candidate-reference differences related to acceptable linguistic variation and the differences related to MT errors.** We devise an alignment-based evaluation metric that compares local syntactic contexts of matching candidate and reference words assessing the impact of (dis-)similarities between human translation and MT output on MT quality, as perceived by human judges. The metric integrates distributional similarity and syntactic equivalence to avoid penalizing acceptable differences, while using context evidence to restrict the amount of admissible variation to the cases where matching candidate and reference words have the same or similar functions in the corresponding sentences. In this part of our work we focus on into-English translation. As mentioned before, we concentrate on sentence-level MT evaluation and use sentence-level correlation with human judgments as the measure of performance of different automatic evaluation strategies.
- **To integrate the fluency aspect of translation quality into reference-based evaluation, as a complementary source of information for detecting acceptable candidate-reference variation.** MT segments that are not matched with the reference translation are assumed to be incorrect. However, as discussed earlier a low number of matches between MT output and a particular human translation is not necessarily indicative of low MT quality. We suggest that explicitly measuring the fluency of the MT segments that are not aligned to any words in the reference mitigates this issue. If an MT segment is fluent, it is more probable that the differences from the reference translation are due to acceptable variation. We integrate fluency information into reference-based evaluation leveraging recent advancements in the use of machine learning techniques to combine different measurements in a single evaluation score. Our approach was ranked as one of the best performing metrics in terms of sentence-level correlation with human judgments at WMT16 Metrics Task (Bojar et al., 2016).

1.4 Thesis Overview

We organized this dissertation following the chronological order of our work. We start by introducing the field of MT evaluation in Chapter 2. In the first part of this Chapter (Section 2.1), we review existing methods for automatic evaluation. The second part (Section 2.2) focuses on manual evaluation.

Chapter 3 is dedicated to the effect of translation shifts on automatic MT evaluation. Theoretical concepts describing the differences between original and translated texts are discussed in Section 3.1, while Section 3.2 presents our experiments on automatic generation of close translation variants for reference-based evaluation.

Chapters 4 and 5 correspond to the central part of our work focused on the improvement of automatic MT evaluation methods. Chapter 4 describes the development of our automatic evaluation metric. Section 4.1 discusses how the relation between candidate and reference words can be properly established through the use of monolingual alignment with syntactic evidence. In Section 4.2 we develop the idea of a syntactic context penalty to be applied to the lexical matches between candidate and reference translations. In Section 4.3 a detailed presentation of the metric is provided. Sections 4.4 and 4.5 contain a detailed account of the metric performance in various evaluation settings.

Chapter 5 introduces our experiments dedicated to the integration of the fluency aspect of translation quality into reference-based evaluation. In Section 5.1 we present a feature-based approach to evaluation describing in detail the adequacy-oriented and fluency-oriented features combined using machine learning techniques. Experimental setup and results are discussed in Sections 5.2 and 5.3.

The last part of our work presented in Chapter 6 is dedicated to the influence of reference translation on manual quality assessment. Section 6.1 presents our experiment on manual evaluation with different reference translations. Section 6.2 provides an analysis of the results obtained. Finally, Chapter 7 summarizes the main conclusions and lists the contributions of this dissertation.

1.5 Publications

Parts of this thesis have appeared previously in the following peer-reviewed publications:

Marina Fomicheva and Núria Bel. Using Contextual Information for Machine Translation Evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2755–2761, 2016.

Marina Fomicheva and Lucia Specia. Reference Bias in Monolingual Machine Translation Evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 77–82, 2016.

Marina Fomicheva, Núria Bel, and Iria da Cunha. Neutralizing the Effect of Translation Shifts on Automatic Machine Translation Evaluation. In *Computational Linguistics and Intelligent Text Processing*, pages 596–607, 2015a.

Marina Fomicheva, Núria Bel, Iria da Cunha, and Anton Malinovskiy. UPF-Cobalt Submission to WMT15 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 373–379, 2015b.

Marina Fomicheva, Núria Bel, Lucia Specia, Iria da Cunha, and Anton Malinovskiy. CobaltF: A Fluent Metric for MT Evaluation. In *Proceedings of the First Conference on Statistical Machine Translation*, volume 2, pages 483–490, 2016.

Chapter 2

MACHINE TRANSLATION EVALUATION

In this Chapter we give a broad introduction to the field of MT evaluation. Evaluation plays a key role in the development of MT systems. Parameters of translation models are optimized to maximize the quality of MT output and the advantage of different approaches is decided based on the results of the assessment of translation quality. Specifically, in the context of system development, MT evaluation serves three main purposes (Gimenez, 2008):

- Error Analysis. Fine-grained evaluation mechanisms are needed in order to detect and analyze MT errors guiding system development and improvement.
- System Comparison. Objective and cost-effective evaluation is necessary to compare different MT systems or different versions of the same system in order to analyze advantages and downsides of different MT strategies or to test how the changes introduced to the system design, training data, etc. affect its performance.
- System Optimization. Modern statistical MT paradigm requires a measure of translation quality for tuning system parameters.

This thesis is primarily dedicated to automatic evaluation. In the first part of this Chapter, we present existing evaluation metrics. We also discuss manual evaluation methods in Section 2.2, since the results of manual evaluation are treated as a ground truth for the assessment of the performance of automatic metrics, and the consistency of manual evaluation is critical for a meaningful comparison between different automatic evaluation strategies.

2.1 Automatic Evaluation

The vast majority of existing evaluation methods are based on the assumption that "the closer a machine translation is to a professional human translation, the better it is" (Pa-

pineni et al. (2002, p.311)). For example, one of the most popular approaches is the so called n-gram-based approach that proceeds by counting the number of matching word n-grams between candidate and reference translations. Due to their cost-efficiency, automatic evaluation metrics allow to considerably accelerate the MT development cycle. They also make possible the use of discriminative training in statistical MT where system parameters are optimized to maximize MT quality measured by an evaluation metric of choice. However, according to Callison-Burch et al. (2007, p.139), "while automatic measures are an invaluable tool for the day-to-day development of machine translation systems, they are imperfect substitute for human assessment of translation quality".

In our view, the main problem of existing reference-based metrics is their inability to reliably distinguish between acceptable and non-acceptable candidate-reference differences. On the one hand, metrics scores are heavily biased by the reference translation. As shown by Lommel (2016), on average there are more differences between alternative human translations of the same source sentence than between an MT output and a human translation. In other words, given a different reference the scores for the same MT output vary more than for translations produced by different MT systems (Culy and Riehemann, 2003). This heavily affects the correlation between the metrics scores and human judgments. Different translation choices found in the MT and the reference that in many cases are considered equally valid by human judges (however, see discussion in Chapter 6), tend to be harshly penalized by automatic evaluation metrics. On the other hand, as we will see in what follows, non-acceptable candidate-reference differences resulting from serious MT errors are not properly penalized by evaluation metrics if the MT output contains a high number of local lexical matches with the reference.

After presenting in detail three of the most widely used evaluation metrics, BLUE, TER and Meteor, in Section 2.1.1, Sections 2.1.2 and 2.1.3 discuss various methods developed to address these two sides of the problem. Furthermore, we review existing alternatives to reference-based evaluation, where features extracted from the source sentences, MT outputs and general target language corpora are commonly used to predict translation quality. Finally, we review approaches based on metric combination which have recently gained attention due to substantial improvement in correlation with human judgments over individual metrics.

2.1.1 BLEU, TER, Meteor

The first family of MT evaluation metrics are based on the proportion of matching or similar words in candidate and reference translations. Most of them combine precision, recall and word order information.[1] BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) are precision-oriented measures. ROUGE (Lin and Och, 2004) and CDER

[1] In the context of MT evaluation *precision* is the proportion of matching words out of the total number of words in the candidate translation, and *recall* is the proportion of matched words out of the total number of words in the reference translation. In other words, precision measures how many words in the

(Leusch et al., 2006) focus on recall. Metrics such as GTM (Melamed et al., 2003) and Meteor (Banerjee and Lavie, 2005; Denkowski and Lavie, 2010a, 2014) calculate similarity balancing recall and precision. Here we describe in detail the well known evaluation metrics used by default in MT evaluation campaigns and as benchmarks in MT evaluation experiments: TER (Snover et al., 2006), BLEU (Papineni et al., 2002) and Meteor (Banerjee and Lavie, 2005; Denkowski and Lavie, 2010a, 2014). For a general review of n-gram-based metrics the reader is referred to Giménez (2008).

One of the first metrics used for the purposes of automatic MT evaluation, Word Error Rate (WER) (Nießen et al., 2000), was borrowed from the field of speech recognition. WER is based on the Levenshtein distance (Levenshtein, 1966), defined as the minimum number of editing steps – insertions, deletions and substitutions – necessary to match two sequences, which is found using dynamic programming approach. Given the Levenshtein distance, WER is computed as the total number of edits normalized by the length of the reference translation. The GALE (Global Autonomous Language Exploitation) (Olive, 2005) research program introduced a modification of this measure called Translation Error Rate (TER) (Snover et al., 2006) that includes a shift operation which moves a contiguous sequence of words within the candidate translation. All edits, including shifts of any number of words, by any distance, have equal cost. More advanced metrics based on edit distance, learn different costs for different edit operations and allow for synonym and paraphrase matching (Snover et al., 2009b).

One of the most widely used evaluation metrics, BLEU (Papineni et al., 2002) was developed to address some of the downsides of WER. BLEU was specifically designed to be used with multiple references in order to allow for acceptable differences in word order and word choice.[2] BLEU scores are calculated as the geometric mean of n-gram precision (with maximum n-gram length of 4) multiplied by a brevity penalty. Precision is measured by dividing the number of n-grams of MT that appear in one of the references by the total number of n-grams of MT. N-grams can be matched against any of the available references. This allows for acceptable variation, but makes it difficult to capture recall, i.e. the number of reference words or word sequences found in the MT output. To see why this is problematic, consider as an extreme example that an MT output containing only one word present in the reference would obtain the unigram precision of $1/1$. In order to avoid this, BLEU uses a brevity penalty that penalizes MT outputs that are shorter than the reference translation.

Formally, BLEU score is defined as:

MT output are correct (also occur in the reference translation), whereas recall computes how many of the correct words were generated by the MT system (i.e. how many of the reference words occur in the MT output). MT can leave out important material and can also add superfluous words, therefore, these measures are complementary. They are typically combined and weighted by means of the F-measure.

[2]It must be noted, however, that in practice, multiple human translations for the same sentences are rarely available.

$$BLEU = BP \times \left(\prod_{n=1}^{4} p_n \right)^{\frac{1}{4}} \tag{2.1}$$

where p_n measures the modified n-gram precision between a document with candidate translations and one or various human reference documents, and BP (Brevity Penalty) down-scales the score for outputs shorter than the reference. The modified precision is defined as:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{\text{n-gram} \in C} Count_{clip}(\text{n-gram})}{\sum_{C' \in \{Candidates\}} \sum_{\text{n-gram}' \in C'} Count_{clip}(\text{n-gram}')} \tag{2.2}$$

where $Candidates$ stands for the set of candidate sentences to be evaluated, $Count$ (n-gram) counts the number of times the n-gram appears in the candidate translation, and $Count_{clip}$ (n-gram) is the same albeit clipped so that it does not exceed the number of times it appears in one of the reference sentences. Brevity penalty is computed as:

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp(1 - \frac{r}{c}) & \text{if } c \leq r \end{cases} \tag{2.3}$$

where c is the aggregate length of the candidate sentences and r is the length of the reference.

Evaluation is performed on a sentence by sentence basis. Then, in order to obtain system level score, BLEU sums up the results for each sentence and divides it by the total number of sentences. BLEU was initially designed for system-level evaluation, i.e. a single metric score generated for a set of translated sentences produced by an MT system. According to Papineni et al. (2002, p.318) "BLEU's strength is that it correlates highly with human judgments by averaging out individual sentence judgment errors over a test corpus rather than attempting to divine the exact human judgment for every sentence: *quantity leads to quality*". Indeed, early studies report a high correlation with human judges at system level. An early study by Coughlin (2003) presents a large-scale investigation involving multiple language pairs and MT systems of the relation between BLEU scores and human judgments. The study reports a high correlation with human judgments averaged over several hundreds of sentences of varying quality (Coughlin, 2003).

In a short document or a sentence, there is a high probability of obtaining zero precision for higher order n-gram counts, which makes the overall BLEU score equal to zero due to the use of geometric mean. To adapt BLEU for sentence-level evaluation, the standard approach is to use smoothing techniques. See (Chen and Cherry, 2014) for a systematic comparison of existing smoothing methods. The most common approach

used in NIST official BLEU toolkit `mteval-v13a.pl` assigns a geometric sequence starting from $1/2$ to the n-grams with zero counts.

BLEU has been used by default for system comparison and parameter optimization. Despite its wide use and practical utility, n-gram-based metrics have been severely criticized. Numerous studies dedicated to the analysis of the behavior of n-gram-based metrics (Culy and Riehemann, 2003; Callison-Burch and Osborne, 2006; Koehn and Monz, 2006) have shown that although BLEU correlates fairly well at system level, even the smoothed version of BLEU is not reliable for assessing the quality of individual sentences.

In a seminal paper, Callison-Burch and Osborne (2006) showed that an improved BLEU score is neither necessary nor sufficient for achieving an actual improvement in translation quality. Furthermore, they demonstrated that BLEU underestimates the quality of MT produced by rule-based MT systems in favor of statistical MT. Similar observations were reported by Och et al. (2003) and Charniak et al. (2003) regarding the performance of statistical MT enhanced with syntactic features that attained a significantly improved translation quality according to manual evaluation, but BLEU reflected a drop in translation quality.

The underlying reason behind this is that BLEU makes independence assumptions that are very similar to the ones made by statistical MT systems. Standard statistical MT models do not have any explicit knowledge of the linguistic structure of the text. When generating a translation, they proceed locally, exhaustively exploring a small context window around the current position and selecting the option that seems optimal given a set of models. This often results in translations that contain correct phrases stacked together in a wrong order with some words added or omitted spuriously, which in the best case are disfluent but understandable and in the worst case completely gibberish. BLEU counts local candidate-reference matches producing relatively high scores for the sentences containing this kind of errors, whereas similar or even lower scores are assigned to the sentences that contain a small number of matches with the reference due to the use of different but semantically equivalent constructions.

Callison-Burch and Osborne (2006) mention the following specific limitations of n-gram-based metrics that guided research in reference MT evaluation in the following years:

- No explicit constraints are placed on the order in which the matching n-grams occur.
- All words matches are equally weighted.
- Only exact matches between candidate and reference words are allowed.
- Recall is not reflected appropriately.

Meteor (Banerjee and Lavie, 2005; Denkowski and Lavie, 2010a, 2014), another well known evaluation metric, was specifically designed to improve on the above limitations of BLEU and achieve a reasonable correlation with human judgments at sentence

level. To address the poor handling of recall, Meteor score is computed as an F-measure taking both precision and recall into consideration. Furthermore, word order differences between candidate and reference translations are explicitly penalized. Also, different values are assigned to content and function word matches to put more weight on the handling of content-bearing material. Finally, in addition to exact word matches, stem, synonym and paraphrase matches are allowed (see next Section for a detailed discussion of the use of lexical-semantic resources in MT evaluation).

Specifically, Meteor's sentence-level score ($Score$) is based on a weighted combination (F_{mean}) of Precision (P) and Recall (R) over different types of lexical matches (m) (exact, stem, synonym and paraphrases) between candidate (h) and reference (r) words. Differences in word order are explicitly penalized by a fragmentation penalty (Pen).

$$Score = (1 - Pen) \times F_{mean} \tag{2.4}$$

$$F_{mean} = \frac{P \times R}{\alpha \times P + (1 - \alpha) \times R} \tag{2.5}$$

$$P = \frac{\sum (\delta \times m_i(h_c) + (1 - \delta) \times m_i(h_f))}{\delta \times |h_c| + (1 - \delta) \times |h_f|} \tag{2.6}$$

$$R = \frac{\sum (\delta \times m_i(r_c) + (1 - \delta) \times m_i(r_f))}{\delta \times |r_c| + (1 - \delta) \times |r_f|} \tag{2.7}$$

Fragmentation penalty is computed as follows. First, the sequence of matched unigrams between the two strings is divided into the smallest number of chunks such that matched unigrams in each chunk are adjacent and identically ordered. The counts of chunks (ch) and matches (m) are then used to calculate a fragmentation fraction: $frag = ch/m$. The penalty is then computed as:

$$Pen = \gamma \times frag^{\beta} \tag{2.8}$$

The weights for different types of lexical matches, as well as the parameters controlling the weights of function and content words (δ), the weights of precision and recall (α) and the parameters controlling the shape of fragmentation penalty function (γ and β) can be optimized to maximize the correlation with human judgments (Denkowski and Lavie, 2010a).

Meteor is different from n-gram based metrics in the sense that instead of directly counting the number of matches between the n-grams of different lengths, an alignment between candidate and reference words is first established, and the score is then computed based on the properties of word correspondences (type of lexical matches and number of sequences of adjacent matching words). Candidate-reference alignment is produced as follows. First, a search space of all possible matches between the candidate and reference words is built by applying all the matchers in sequence. Once all possible

matches have been identified, the alignment is produced as the largest subset of these mappings. If more than one alignment is possible, the alignment which best preserves word order (i.e. contains the fewest "crossing" unigram mappings) is selected. We note that alignment-based formulation is more informative than n-gram matching, precisely because it allows to take into account the properties of (dis-)similarities between candidate and reference translations. In Section 4.1 we discuss in detail the advantages of the alignment-based approach to evaluation.

Meteor outperforms both TER and BLEU by a large margin (see Appendix A for the results of various evaluation campaigns), on account of addressing acceptable variation in translation, an explicit penalization of word order differences and weighting the differences based on the distinction between content and function words. In the next Sections we discuss in more detail the work dedicated to the problem of acceptable vs. non-acceptable variation in reference-based evaluation.

2.1.2 Acceptable Variation

As we have seen in the previous Section, an important family of evaluation methods is based on surface-string matching between MT and human reference translation. All of them compare candidate MT to one (or various) human references and compute an evaluation score based on the number of matching words or word sequences.

Translation is an open task with multiple possible solutions. Therefore, the fact that an MT output does not contain the exact same words as the available reference translation, does not necessarily indicate that it contains translation errors. Below we describe existing proposals that aim at increasing the coverage of acceptable linguistic variation between candidate and reference translations.

To allow for morphological variation, character n-grams can be used instead of word n-grams (Yang et al., 2013; Stanojevic and Sima'an, 2014; Popovic, 2015). Consider the following example adopted from (Yang et al., 2013): A metric based on word n-

MT: Tom has an interest in cooking.
Ref: Tom is interested in cooking.

grams would not be able to relate the words "interest" and "interested", although here they clearly contribute to the sentence similarity between candidate and reference translations. By contrast, a metric that uses character n-grams as the units of comparison would correctly establish a match between the character n-grams that constitute the stem of these words.

Yang et al. (2013) were the first to use character n-grams for MT evaluation. They substitute the word by the character as a unit of comparison in n-gram-based evaluation metrics and achieve an improvement over some of the word-based metrics. Popovic

(2015) designed the ChrF evaluation metric based on character n-grams F-score. The method has shown very promising results achieving the best performance for out-of-English translation at the WMT14 Metrics Shared Task (Macháček and Bojar, 2014b). Stanojevic and Sima'an (2014) utilize character n-grams as one of the features in their feature-based evaluation metric BEER, which outperforms most of the participating systems at WMT14 and WMT15 Metrics Tasks (Macháček and Bojar, 2014b; Stanojevic et al., 2015), using no additional linguistic processing resources. An alternative way to account for morphological variation is to use a more abstract word representation, such as lemmas or stems. For example, Meteor employs Porter stemming (Porter, 1980) to align different words that share the same stem, i.e. they are morphological variants of each other.

Variation between possible translations goes far beyond the morphological level. Different translations of the same source sentence may contain synonyms. Words with similar meaning can be detected using existing lexical resources. For example, Banerjee and Lavie (2005) in Meteor were the first to use WordNet synsets in order to allow for synonym matches. WordNet (Fellbaum, 1998) is a large lexical database that contains nouns, verbs, adjectives and adverbs organized in sets of synonyms called synsets, each expressing a specific concept. Synsets are linked to each other according to semantic, conceptual and lexical relationships. Polysemous words are included in multiple synsets. Princeton WordNet, a lexical resource for English, contains more than 117,000 synsets. Other lexical-semantic resources are also available for this purpose, e.g. BabelNet (Navigli and Ponzetto, 2010) or DBnary (Sérasset, 2015).

However, it is not always possible to establish semantic equivalence at word level. In order to match similar phrases, it has been proposed to use a pre-computed paraphrase tables, typically obtained with the pivot phrase method (Bannard and Callison-Burch, 2005). This method employs various aspects of phrase-based statistical MT including phrase extraction heuristics to obtain bilingual phrase pairs from word alignments. To obtain monolingual paraphrases, first, a single bilingual phrase table is built for a language pair from a parallel dataset. The source phrases that translate into the same target phrase are considered paraphrases of each other. Multiple paraphrases are frequently extracted for each phrase and can be ranked using a paraphrase probability based on phrase translation probabilities.

As discussed in Callison-Burch (2008), the problems involved in bilingual phrase extraction familiar from statistical MT are propagated to the paraphrases obtained with the pivot method. Specifically, the heuristics used in statistical MT for the extraction of bilingual phrases from word alignments allows unaligned words to be included at the boundaries of the source and target phrases. This populates the resulting paraphrase table with noisy phrase pairs. Various filtering techniques have been proposed to address this issue including filtering out the paraphrase pairs that were obtained from bilingual phrases with low probability (Snover et al., 2009a; Denkowski and Lavie, 2010a) or

using syntactic constraints (Callison-Burch, 2008). Various paraphrase resources are freely available either independently or as part of other Natural Language Processing (NLP) systems (Callison-Burch, 2008; Snover et al., 2009b; Ganitkevitch et al., 2013; Denkowski and Lavie, 2014).

Although even after filtering paraphrase tables contain a considerable amount of noisy co-occurrences that are completely unrelated, various MT evaluation metrics successfully use synonym and paraphrase support either to enhance candidate-reference alignment (ParaEval (Zhou et al., 2006), Meteor (Denkowski and Lavie, 2010a), TER-Plus (Snover et al., 2009b)), or to generate paraphrases of the human translation to be used as additional references (Kauchak and Barzilay, 2006; Owczarzak et al., 2006; Madnani et al., 2007). In this thesis we explore both of these possibilities in Chapters 3 and 4, respectively.

2.1.3 Non-acceptable differences

As previously suggested, one of the main problems of reference-based evaluation resides in the inability of the metrics to reliably distinguish between acceptable and non-acceptable candidate-reference differences. One side of this problem discussed in the previous Section is that candidate-reference differences deemed acceptable by human judges can drastically decrease automatic evaluation scores. The other side of the problem is that true divergences stemming from serious MT errors are not properly penalized by evaluation metrics if the number of local lexical matches is high.

Liu and Gildea (2007) proposed to use source sentence information in order to avoid an increase in n-gram-based evaluation scores produced by spurious matches between candidate and reference words that correspond to different parts of the source sentence. Both candidate and reference translation are aligned to the source sentence using bilingual word alignment and then n-gram matches are computed only for the words that are aligned to the same words in the original sentence. Source-constrained n-gram precision calculated in this way outperforms simple n-gram matching strategies by a significant margin. In Chapter 4 we propose an alternative solution to the problem of spurious n-gram matches that does not require the use of noisy bilingual alignment.

In addition, some of the approaches already discussed above partially address this problem. For instance, the introduction of an explicit penalty for word order alterations in Meteor restricts the amount of admissible variations between candidate and reference translations. The use of additional linguistic information, at syntactic (Mehay and Brew, 2006; Owczarzak et al., 2007; Habash and Elkholy, 2008; Giménez, 2008; Kahn et al., 2009; Popović and Ney, 2009; He et al., 2010) and semantic levels (Reeder et al., 2001; Giménez, 2008; Macháček and Bojar, 2011; Lo et al., 2012) is another strategy to ameliorate this issue.

Abstract linguistic representation allows, on the one hand, to abstract away from surface word forms and sequential word order (thus addressing the issue of acceptable

differences discussed in the previous Section), and on the other hand, to assess the grammaticality of the MT output, penalizing translations that contain roughly the same words but are grammatically ill-formed.

One of the most influential works on using syntactic information for MT evaluation was presented by Liu and Gildea (2005). They developed several measures of syntactic similarity. The measures are derived from BLEU, but instead of word n-grams, fractions of syntactic trees are employed. Liu and Gildea (2005) use two types of syntactic representation: constituent and dependency trees. Specifically, they developed a Syntactic Tree Matching (STM) metric based on constituency parsing, and a Head-Word Chain Matching (HWCM) metric based on dependency parsing. STM computes the number of matching fractions of syntactic subtrees with different depths and their arithmetic mean is used as the evaluation score. HWCM is based on the number of matching head-word chains extracted from a dependency-based representation. A head-word chain is defined as a sequence of words that corresponds to a path in dependency tree. The evaluation score is computed in a manner similar to BLEU but using n-grams of dependency chains instead of n-grams in the linear order of the sentence. The performance of the metrics is tested using human judgments of fluency and of overall translation quality on a 5-point scale. The results show that syntactic measures outperform BLEU with HWCM achieving a more stable performance. The reason probably is that STM completely abstracts away from the string-level representation. Using this measure on its own is problematic, as it completely ignores word choice.

Giménez and Màrquez (2010) developed a general framework, where the overlap between candidate and reference linguistic elements is computed as a quality measure. A linguistic element is defined as an abstract reference to any possible type of linguistic unit, structure or relation. Thus, sentences can be represented as sets of words, syntactic constituents, named entities, semantic roles or discourse representations. Giménez and Màrquez (2010) explore different configurations of the proposed representations: lexical overlap over syntactic constituents, semantic roles, named entities, overlap between semantic roles, part-of-speech overlap over discourse representations. The results show that linguistic measures are outperformed by lexical measures in terms of sentence-level Pearson correlation.

Intuitively, comparing high-level abstract representations is advantageous for reference-based evaluation, as it provides an abstract representation which highlights meaning-related aspects of quality while allowing certain variation at surface lexical and syntactic levels. However, the noise introduced by automatic linguistic analysis hinders the benefits of using such representations. Furthermore, while such methods may work well for sentence similarity tasks (e.g. paraphrase detection or textual entailment recognition), MT evaluation differs from those in a very important way. MT outputs contain errors that affect the perception of quality to a varying extent. While a lack of overlap between semantic roles shows that the meaning of candidate and reference sentences is

different, it is not indicative of how bad the translation actually is. Finally, linguistically informed metrics typically measure different linguistic aspects (e.g. lexical and syntactic) separately, combining the measurements only at sentence level. In our view, this is not optimal. To assess the quality of a translation, the quality of its constituent parts needs to be assessed. Assessing the correctness of a word or an expression involves both lexical and syntactic aspects. In Section 4.3 we will show how to combine this information at a lower level of granularity in order to obtain a better estimate of sentence-level quality.

As a final remark in this Section, we must mention the work dedicated to automatic error analysis. This is a very promising but still rather unexplored area. Among the existing work we find for example the framework for automatic error analysis and classification developed by Popović and Ney (2011). Their proposal is based on the identification of actual erroneous words using the algorithms for computation of Word Error Rate (WER) (see Section 2.1.1) and Position-independent Word Error Rate (PER). Popović and Ney (2011) worked with five error categories: inflectional errors, errors related to wrong word order, missing words, extra words, and incorrect lexical choices. The extracted erroneous words were then used in combination with different types of linguistic knowledge (e.g. part-of-speech tags, named entity tags, compound words, etc.) in order to obtain various details about the nature of actual errors.

Fishel et al. (2012) applied this idea for the ranking evaluation task. They compared MT outputs based on the frequencies of different error categories.[3] The task was framed as a binary classification problem where the classifier must assign a win to one of the two MT outputs under comparison. The systems showed very modest results in terms of the correlation with human preferences at sentence level due to the problem of sparsity, i.e. many zero counts of specific error types.

2.1.4 Alternatives to Reference-based Evaluation

So far we have discussed some of the important limitations of reference-based evaluation, as well as various strategies intended to amend them. Many issues in MT evaluation arise from the fact that whereas many translations of the same sentence are in principle possible, only one reference is usually available to be used for evaluation. In this Section we review the work dedicated to the automatic generation of additional references, as well as reference-free evaluation methods, where the use of reference translation is avoided altogether.

[3] As will be suggested in Section 4.3, besides the type of error (e.g. morphological, lexical, word order), the function and position of the affected words in the sentence can have a varying impact on the perceived quality of the MT.

Automatic Generation of Additional Reference Translations

As mentioned before, lexical resources can be used not only to enhance candidate-reference comparison, but also to generate additional reference translations in order to avoid penalizing the MT outputs that are essentially correct but differ from the available reference translation:

> ... a candidate translation expressing the source meaning accurately and fluently will be given a low score if the lexical choices and syntactic structure it contains, even though perfectly legitimate, are not present in at least one of the references. Necessarily, this score would not reflect a much more favorable human judgment that such a translation would receive (Owczarzak et al. (2006, p.86)).

Studies using paraphrase generation to obtain additional references (Kauchak and Barzilay, 2006; Owczarzak et al., 2006; Madnani et al., 2007; Albrecht and Hwa, 2007; Madnani and Dorr, 2013) have demonstrated that enriching the space of possible translation solutions, even with noisy references, tends to improve the performance of the metrics.

An interesting observation made by Snover et al. (2006) is that some reference translations may be more suitable as the point of comparison than others: "[...] a more successful approach [to evaluation] is one that finds the closest possible reference to the hypothesis from the space of all possible fluent references that have the same meaning as the original references" (Snover et al. (2006, p.226)). Based on this idea, Snover et al. (2006) propose a semi-automatic measure of translation quality called HTER (Human-targeted Translation Edit Rate). Recall from Section 2.1 that TER measures the minimum number of edits necessary to correct the MT output, so that it conveys the same meaning as a reference translation. To obtain the so called "targeted" reference, fluent speakers of the target language are asked to produce new translations starting with the MT output and one or more untargeted reference translations. It is explicitly indicated to the annotators that the new translation must minimize TER (i.e. to be as close as possible to the MT output). Annotators can generate the targeted reference by editing the system output or the original reference translation. Snover et al. (2006) reported experiments using 200 MT sentences evaluated for adequacy and fluency on a 5-point scale from the MTEval 2004 Arabic evaluation dataset. Automatic evaluation with targeted reference outperforms standard reference-based metrics by a very large margin. Interestingly, HTER correlates better with average human judgments than human judgments coming from different annotators correlate with each other.

HTER is widely used in MT evaluation campaigns. In practice, the data for computing this metric is produced from MT post-editing and its use is mainly limited to computer-assisted translation scenarios. Madnani and Dorr (2013) developed a method for automatic generation of targeted reference translations for parameter tuning of statistical MT. Following their work in Madnani et al. (2007), where the task of paraphrase

generation was addressed as an English–English MT problem, Madnani and Dorr (2013) designed a method that uses both the initial reference translation and the MT output for the generation of additional paraphrases (thus resembling the process of manual generation of targeted reference described above). They report improvements in MT performance when conducting parameter tuning using the additional automatically generated references.

Reference-free Evaluation

Instead of human reference translation, Albrecht and Hwa (2007) use pseudo-references, i.e. MT outputs generated by different MT systems. Pseudo-references are not perfect translations. Even if the MT under evaluation were identical to a pseudo-reference, i.e. another MT system output, it might not be a good translation. An interesting shift in perspective in this work is to view the reference not as an ideal to strive for but as a benchmark to compare against.

Albrecht and Hwa (2007) propose a feature-based approach that combines different measurements in a single score assigning different weights obtained through machine learning techniques to each measurement (see Section 2.1.5). Their evaluation model is able to learn how much to trust the pseudo-references generated by certain MT system, as the reference MT systems are calibrated against an existing set of human judgments for the outputs of the MT system under evaluation.

The results show that metrics do significantly better with four pseudo-references than with one human reference. In addition, they do almost as well with four machine generated references as when using four human references. A metric that compares MT output against a diverse population of differently imperfect sentences is more discriminative in judging MT quality than only comparing against gold standards.

An alternative approach is to avoid using the reference translation altogether and extract information for the evaluation from the MT output itself, from the source sentence, from the relation between the two or from large target language corpora. This approach has been widely explored in the task called quality estimation[4] which aims at predicting the quality of the output of MT systems in use and therefore does not have access to any reference translation (Specia et al., 2010b). Applications of quality estimation include deciding which segments need to be revised by a professional translator, deciding whether a reader gets a reliable gist of the text and estimating how much effort it will take to post-edit a segment or select among alternative translations produced by different MT systems.

Reference-based MT evaluation and quality estimation have different use scenarios but pursue the same goal of accurately predicting translation quality (although the

[4]Initially, the task was conceived as a confidence estimation problem, aiming to estimate the confidence of the MT system in the generated output. Quality estimation is a more general term referring to reference-free evaluation with a wide range of features including those independent from the MT system.

definitions of quality may vary depending on the intended use of MT). However, these two tasks have been developing in parallel, with a few notable exceptions attempting to leverage the achievements of the work accomplished in both fields. In this work, we take advantage of the advancements in quality estimation task (see Chapter 5). Below we describe the basic methods used in reference-free evaluation.

Quality estimation was initially viewed as a binary classification problem of distinguishing between "good" and "bad" translations (Blatz et al., 2004). Later the task was extended to predicting discrete or continuous scores in a given range. The gold standard scores can be obtained either using reference-based metrics (Blatz et al., 2004) or manual evaluation (Quirk, 2004). Human judgments against which quality estimation is compared are traditionally based on post-editing effort (e.g. post-editing time, HTER scores or a subjective score reflecting how much effort is required to post-edit the MT output). Besides sentence-level quality, quality estimation may be addressed at word level. In the latter case, each word in the MT output is judged as correct or incorrect, or labeled for a specific error type (Bach et al., 2011; Luong et al., 2015). Word-level systems are typically trained and evaluated using post-editing data, where each MT word is assigned a "bad"/"good" label depending on whether it was changed during the post-editing process.

Quality estimation is thus typically addressed as a supervised machine learning task where the goal is to predict some quality labels based on a set of features extracted using source sentences, MT outputs, internal MT system information and source and target language corpora. The work on quality estimation has largely focused on the design of features and the selection of appropriate learning algorithms. Broadly, the features developed for quality estimation can be divided in *glass-box* and *black-box*.

Glass-box features exploit the information on the decoding process of the MT system under evaluation, such as hypothesis scores, n-best lists and phrase probabilities. Black-box features are system independent. They are extracted from the source and target texts, using additional resources and large source and target corpora. The interest in the design of black-box features has shifted the focus from estimating the confidence of statistical MT to a more general task of evaluating MT quality.

The principal features that have been proposed in the literature for sentence-level quality estimation are intended to capture:

(a) relation between the source and target sentences (e.g. ratio of the source and target sentence lengths, statistical MT alignment score, percentage of different types of word alignments)[5],
(b) difficulty of the source sentence (e.g. length of the source sentence, average source word length, percentage of 1 to 3-grams in the source sentence belonging

[5]Conceptually, this is close to human evaluation. However, to be able to accurately compare the meaning of the source and target sentences one would need to have solved the MT problem.

to each frequency quartile of a monolingual corpus, average number of translations per source word in the sentence (as given by probabilistic dictionaries)),

(c) fluency of the target sentence (e.g. language model log-probabilities and perplexities.

The most popular choice for the algorithm for training sentence-level quality estimation models is a regression or classification implementation of Support Vector Machines (SVM) with Radial Basis Function (RBF) kernel. Word-level quality estimation is traditionally treated as a structured prediction problem and addressed using Conditional Random Fields (CRF).

Sentence-level quality estimation models have been reported to beat baseline reference-based systems. For example, Specia et al. (2010b) report a significant improvement attained by quality estimation systems over BLEU, NIST, TER and Meteor in terms of sentence-level Pearson correlation with human judgments based on post-editing criterion, on EAMT09 English-Spanish dataset consisting of 4,000 sentences translated by four statistical MT systems (Specia et al., 2009). However, quality estimation performance varies considerably depending on the language pair, MT systems and the quality score being optimized.

Features that measure the probability of the MT output with respect to a target Language Model (LM) have been shown to contribute strongly to the performance of quality estimation models both for sentence-level and word-level tasks. The intuition behind these features is that a perplexity score from an LM trained on an in-domain target language corpus will reflect the degree to which the observed word sequence is expected compared to what has been observed in the training corpus, thus capturing the fluency of the MT output. In this work we explore this idea for the integration of the fluency aspect in the reference-based evaluation (see Chapter 5).

2.1.5 Metric Combination and Feature-based Approaches

So far we have presented various ways for computing the similarity between candidate and reference translations based on different representations, as well as the use of alternative sources of information, such as original sentences or monolingual corpora. Different aspects of candidate-reference similarity can provide complementary information. An interesting question then is how these pieces of information can be combined in a single score. Below we first discuss different measurements and metrics that have been applied in combination. Next, we present different strategies that have been used for metric combination.

First, note that some of the metrics discussed in this Chapter already combine various different measurements in a single score. For instance, Meteor combines lexical similarity values with a fragmentation penalty which is based on word order differences. Second, various evaluation frameworks combine measurements based on differ-

ent linguistic aspects of candidate-reference similarity (Giménez and Màrquez, 2010; Comelles et al., 2012; Guzmán et al., 2014; Yu et al., 2015). This approach was initially proposed by Giménez and Màrquez (2010). Their underlying motivation was to capture different aspects of translation quality in order to avoid the bias introduced by n-gram-based metrics that focus on lexical similarity. In our view, the idea that comparing syntactic structures captures the syntactic aspect of translation, whereas comparing word n-grams captures the lexical aspect of translation is not quite accurate. Syntactic structure emerges from the use of different word forms, function words and word order. In so far as n-gram-based metrics account for the above characteristics, they also capture the differences in syntactic structure to a certain extent. Metrics based on different linguistic representations (words, syntax, semantics or discourse) differ not in the aspects of translation quality that they measure, but rather in the level of generalization.

Finally, some metrics combine reference-based and reference-free evaluation. Albrecht and Hwa (2007) proposed a regression-based approach to combine metrics with and without human reference. They used four kinds of features to train their regression-based model: a) string-based reference-based metrics; b) syntax-based reference-based metrics (HWCM and STM); c) reference-free metrics using n-gram matching against a target language corpus (similar to LM features from quality estimation framework); and d) reference-free metrics using syntactic sub-trees matching against a target language corpus.

Specia and Giménez (2010) explore the combination of a large set of quality estimation features extracted from the source sentence and the candidate translation, as well as the source-candidate alignment information, with a set of 52 MT evaluation metrics from the Asiya Toolkit (Giménez and Màrquez, 2010). They report a significant improvement over the reference-based evaluation metrics on the task of predicting human post-editing effort.

In order to combine different metrics or measurements, an average of the scores generated by the metrics can be used as the final score. For example, Giménez and Màrquez (2010) use a normalized arithmetic mean of individual metrics. However, different metrics can contribute to the overall score in different ways. Unlike human translation, MT use is still typically limited to specific scenarios (such as post-editing or gisting). Depending on the task, evaluation criteria and results may vary considerably. For example, an omission of a negative particle completely changes the meaning of a sentence, and therefore such translation would obtain a very low adequacy score. By contrast, if the evaluation task consists in predicting post-editing effort, the translation would be judged as requiring very little editing.

Some evaluation metrics, such as Meteor or TER-plus combine various internal components allowing for optimization of a small number of parameters using greedy search. Interestingly, depending on the type of human evaluation different components receive substantially different weights. Denkowski and Lavie (2010b) tuned Meteor pa-

rameters (see Section 2.1.1) using absolute quality judgments, ranking judgments and HTER. First, they found that parameters for HTER obtained for different datasets are the most stable, while parameters for absolute scoring and ranking fluctuate the most. Second, they found that all parameter sets favor recall over precision. Denkowski and Lavie (2010a) suggest various possible reasons for that. First, statistical MT is typically optimized using a precision-based metric (BLEU), which negatively affects recall. Another possible reason is related to how the human evaluation task is formulated. Since annotators are asked how much of the meaning of the human translation is preserved in the MT output, they would first read the reference translation and then see how many of the reference words are also present in the MT output, which corresponds to recall in automatic evaluation. Finally, regarding the fragmentation penalty that captures differences in word order, the ranking task has the slightest fragmentation penalty, reflecting possibly that word order differences as measured by Meteor are not relevant for comparing different MT outputs, followed by the adequacy task, whereas HTER task has the harshest penalty, reflecting the strict requirement that each reordering requires an edit. Snover et al. (2009a) conduct similar experiments optimizing the weights for different edit operations using different types of human judgments including adequacy, fluency and HTER arriving to similar conclusions regarding the impact of word order (shift operations) and precision vs. recall (deletion vs. insertion operations). Thus, different types of human evaluation indeed highlight different characteristics of MT.

Optimizing the weights through greedy search allows to work with a very limited number of features. Recent approaches to metric combination rely on machine learning techniques to alleviate this issue. In this framework, the input (i.e., the MT sentence to be evaluated) is represented as a set of features. The features are separate measurements computed for the candidate-reference pair or individual metric scores. The learning algorithm then tries to find a mapping from input features to a score that quantifies the MT quality by optimizing the model to match human judgments on training examples.

Depending on the type of human judgments (see Section 2.2.3), several learning paradigms have been applied for automatic evaluation:

- Binary functions that classify whether the input sentence is human-translated or machine-translated (Corston-Oliver et al. 2001; Kulesza and Shieber 2004)
- Continuous functions that score translation quality of input sentences on an absolute scale (Quirk, 2004; Lita et al., 2005; Albrecht and Hwa, 2007; Liu and Gildea, 2007; Uchimoto et al., 2007)
- Ordinal functions that give ranking preference between multiple translations (Ye et al., 2007; Duh, 2008)

Some recent approaches based on a combination of many different evaluation metrics have proven to be very successful in recent evaluation campaigns (Stanojevic and

Sima'an, 2014; Guzmán et al., 2014; Yu et al., 2015). We use them as benchmarks in our experiments and present them in more detail in the corresponding Chapters (see Sections 4.4 and 5.3).

It must be noted, however, that the gains in performance attained by metric combinations are not necessarily due to a meaningful motivated choice of complementary evaluation strategies or to the fact that selected metrics capture different aspects of quality. Different metrics (or linguistic representations) have different advantages and limitations and their combination is simply more robust than individual metrics.

2.2 Meta-evaluation

The results of human quality assessment serve as the ground truth for automatic evaluation metrics. In this Section we first review the main types of manual evaluation and their limitations. Next, we explain how the performance of automatic evaluation metrics is assessed by comparing the results of automatic evaluation with human judgments.

2.2.1 Types of Manual Evaluation

Evaluation of translation quality is a challenging task. Besides the contents of the original text translation choices are guided by numerous extra-linguistic factors such as the intended use of translation, differences between background knowledge of the source and target audiences, language-specific genre conventions, etc. The appropriateness of translator's choices is evaluated in the light of these factors.

Some early methodologies used evaluation frameworks developed for the assessment of human translation quality. For example, the "quality panel" approach from the evaluation program of ARPA(White et al., 1994) [6] involved subjecting MT outputs to a panel of professional, native English-speaking translators of the relevant languages who assessed the quality of MT in terms of its lexical, grammatical, semantic, and stylistic accuracy and fluency. This approach was abandoned as too complex and expensive in favor of more intuitive evaluation tasks.

A simpler alternative that has become the standard for manual MT evaluation is the assessment of translation adequacy and fluency on an multi-point scale (Linguistic Data Consortium, 2005). Adequacy measures how much of the meaning of the source sentence is preserved in the MT output. Fluency refers to the well-formedness of the translation. For the assessment of adequacy, human translation is often used instead of the source text. Such monolingual reference-based evaluation is an attractive practical solution since it does not require bilingual speakers. However, as we have shown in

[6] The ARPA MT Initiative is part of the Human Language Technologies Program of the Advanced Research Projects Agency Software and Intelligent Systems Technology Office.

(Fomicheva and Specia, 2016), human annotators, not unlike automatic evaluation metrics, are strongly biased by the reference translation. This issue is discussed in detail in Chapter 6.

An example of the multi-point adequacy and fluency scales for monolingual evaluation is given in Table 2.1:

- Adequacy: How much of the meaning of the professional human translation is also expressed in the MT.
- Fluency: How do you judge the fluency of this translation.

	Adequacy		Fluency
5:	All	5:	Flawless
4:	Most	4:	Good
3:	Much	3:	Non-native
2:	Little	2:	Disfluent
1:	None	1:	Incomprehensible

Table 2.1: Multi-point Adequacy and Fluency scales

This type of evaluation is based on the defining properties of translation and constitutes a powerful and intuitive instrument for assessing MT quality. Measuring absolute quality on an interval-level scale, however, presents a problem of low inter-annotator agreement. The scale is arbitrary and no precise instructions are given to the annotators. As a result, different judges may assign different scores to the same sentence.

To make the task even more intuitive and elicit more consistent judgments, an alternative setting has been introduced, in which the judges are asked to rank different MTs of the same source sentences in terms of their relative quality, allowing ties (Callison-Burch et al., 2007). This formulation of the task eliminates the need to assess the "goodness" or "badness" of translations in absolute terms in favor of simpler preference judgments. Although ranking has become the predominant approach in the recent years, it has some important downsides. Ranking is problematic for longer sentences containing various errors of different types (Bojar et al., 2011). In these cases the annotators can be quite inconsistent in their preferences[7] or assign a tie considering that the system outputs are roughly equivalent, which provides no information at all regarding the quality of the outputs under comparison (they may be equally good or equally bad). Furthermore, ranking does not indicate the magnitude of the differences in quality, making this type of evaluation less informative. See (Denkowski and Lavie, 2010b; Bojar et al., 2011) for a detailed discussion of the limitations of the ranking approach to evaluation.

[7] As an illustration, consider the work of Farrús et al. (2010) who show that differences in ranking are caused by annotators treating some types of errors as more serious.

Finally, in the case of task-oriented evaluation, the problem of inter-annotator agreement is less severe. Rather than directly eliciting absolute or relative judgments, post-editing tasks attempt to measure the minimum amount of editing required to correct the MT output. This is done by computing the minimum edit distance between the MT output and its post-edited version (Snover et al., 2006), typically using the TER evaluation metric. This strategy is limited to a specific use of MT outputs and suffers from the disadvantages inherited from TER, such as exact word matching and equal weights for different editing operations.

The performance of automatic evaluation metrics varies significantly depending on the type of human judgments and the error metric (Denkowski and Lavie, 2010a). Different types of human judgments pose different challenges to automatic evaluation metrics. Ranking can be more difficult when very similar MTs have to be compared, in which case fine-grained distinctions between different kind of errors have to be made. On the other hand, in the ranking task the scores produced by a metric are not assessed directly. Ranking judgments provide little insight regarding how well the magnitude of the differences in quality between the MTs of different source sentences is reflected in automatic evaluation. In this work, we test the performance of our approach to automatic MT evaluation using data from different evaluation tasks (ranking, absolute scoring with an interval-level scale and absolute scoring with a continuous scale) confirming that our approach is robust and stable in different evaluation settings.

2.2.2 Evaluation Campaigns and Datasets

Most of the experiments presented in this work rely on the data available from the evaluation campaigns of the Workshop on Statistical Machine Translation (WMT) held annually since 2006. The test data at the WMT Translation Task consists of news articles extracted from online sources.[8] Human translations are obtained from a professional translation agency. Participating MT systems are varied and representative of the state-of-the-art.

WMT datasets thus include source texts, human reference translations and the outputs from the participating MT systems, for five language pairs. In manual evaluation annotators are presented with the source sentence, its human translation and the outputs of different MT systems and asked to rank the MT outputs from best to worst. Annotations are collected from volunteers from the participating research teams. For efficiency reasons, annotators are asked to compare the outputs of five MT systems (randomly sampled from the dataset) for each sentence at once and rank them from best to worst. Ties are allowed. From this compact annotation, 10 pairwise preference judgments can

[8]Starting from WMT14, MT outputs and human translations are produced in the same translation direction. This is important, since translated texts have properties that make them different from the texts originally written in the target language. See Section 3.1.2 for discussion.

be extracted for each sentence in a straightforward way. For example, if a judge ranked the outputs of the systems A, B, C, D, E as $A > B > C > D > E$, then $A > B$, $A > C$, $A > D$, $A > E$, etc. It should be noted that neither the absolute value of the ranking, nor the degree of the difference is taken into consideration.

At WMT16 a new type of evaluation was tested consisting in absolute quality judgments based on the adequacy criterion. The judgments were collected following the procedure described in Graham et al. (2015) for all available into-English language pairs. Human assessors were asked how much of the meaning of the reference translation was preserved in the MT output. The evaluation was performed using a 0-100 rating scale. Direct assessment scores were standardized according to an individual annotator's overall mean and standard deviation. Up to 15 assessments were collected for each MT output from different assessors and the results were averaged to obtain the final score. 560 MT segments sampled randomly from the data were annotated by humans for each language pair, resulting in a total of 3,360 segments of into-English translations.

Besides WMT datasets, two other datasets are used in our experiments. In Chapter 3 we use the EAMT09 dataset (Specia et al., 2010a) which consists of 4,000 source sentences in English randomly extracted from Europarl (Koehn, 2005), their corresponding human translations into Spanish and MT outputs produced by 4 statistical MT systems. The Europarl corpus is a multilingual, parallel corpus that has been collected from the proceedings of the European Parliament since 1996 and includes translations from (into) 21 official languages of the European Union. Translation is done by professional translators, who are native speakers of the corresponding target languages (van Halteren, 2008).

The EAMT09 dataset (Specia et al., 2010a) includes manual evaluation scores, collected from professional translators using the post-editing criterion. Specifically, annotators were asked to indicate the amount of editing needed in order to make the MT ready for publishing: 1) requires complete re-translation, 2) a lot of post-editing is needed, 3) little post-editing is needed, 4) fit for purpose. Human assessments were obtained independently from the reference translation, from bilingual speakers, using the source text and the MT output.

In Chapter 4 we use the MTC-P4 Chinese-English dataset produced by the Linguistic Data Consortium (LDC2006T04). This dataset contains 919 source sentences from the news domain, 4 reference translations and MT outputs generated by 10 MT systems. The translations produced by 6 of the systems were assigned quality scores following the Linguistic Data Consortium evaluation guidelines (Linguistic Data Consortium, 2005), based on fluency and adequacy criteria, on a 5-point scale. In total, human assessment is provided for 5,514 MT sentences. For adequacy assessment, the annotators were presented with MT output and a human translation, whereas for fluency assessment only MT output was used.

2.2.3 Meta-Evaluation Techniques

To assess the accuracy of MT evaluation metrics, the scores generated by the metrics are typically compared with human judgments. In the case of absolute quality judgments, the Pearson correlation coefficient (r) (Pearson, 1924) is commonly used.[9]

Following recent work on meta-evaluation (Graham and Baldwin, 2014), in our experiments with human judgments of absolute quality we use the Hotelling-Williams test for dependent correlations (Williams, 1959) to compute the significance of the difference between correlations for different evaluation metrics. Correlations computed for two separate automatic metrics on the same data set are not independent, and for this reason, in order to test the difference in correlation between them, the degree to which the metrics correlate with each other should be taken into account. The Hotelling-Williams test takes this into account. The higher the correlation between the metric scores, the greater the statistical power of the test.

For the ranking task, Kendall rank correlation coefficient (τ) between the metrics and the human ranking is computed as follows:

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|} \tag{2.9}$$

where $Concordant$ is the set of all human comparisons for which a given metric suggests the same order and $Discordant$ is the set of all human comparisons for which a given metric disagrees. Specifically, we use the definition of Kendall τ presented in Macháček and Bojar (2014b). This was the official measure for the WMT Metrics Task starting from 2014. In this definition, human ties are ignored, while metrics ties are counted in the denominator.

To assess the significance of the results of the ranking tasks we include empirical confidence intervals, computed using the bootstrap resampling method as proposed in Macháček and Bojar (2014b). However, we must note that in the case of the WMT data presented in the previous Section, Kendall τ is computed in a non-standard way (we do not have a single overall ranking of translations, but rather rankings of sets of 5 translations). As a consequence, it has recently been suggested that the accuracy of confidence intervals computed in this way is difficult to verify (Bojar et al., 2016).

[9] In some work, the Spearman rank-correlation coefficient (Spearman, 1904) is used instead (Albrecht and Hwa, 2007). By contrast to Pearson r that assumes normal distribution, Spearman correlation is a distribution-free test. Albrecht and Hwa (2007) suggest that this is more appropriate given that metrics' scores are not normally distributed. We agree with this observation, but we will use Pearson correlation in our experiments to make the results comparable to existing work.

2.2.4 Inter-Annotator Agreement

Manual evaluation of MT serves both to assess the quality of MT system outputs and to evaluate and/or train automatic evaluation metrics. Therefore, it is very important that human annotation is reliable and consistent. To establish this, the common practice is to have the same data points annotated by multiple annotators and to measure the agreement between them. Inter-annotator agreement is typically computed using the Kappa coefficient (Cohen, 1960):

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.10}$$

where $P(A)$ is the proportion of times annotators agree and $P(E)$ is the proportion of times annotators are expected to agree by chance. κ has a value of at most 1, with higher κ values indicating higher rates of agreement. The exact interpretation of the kappa coefficient is a difficult issue, but according to Landis and Koch (1977), 0-0.2 is slight, 0.2-0.4 is fair, 0.4-0.6 is moderate, 0.6-0.8 is substantial, and 0.8-1.0 is almost perfect.

In the context of pairwise comparisons of MT system outputs, an agreement between two annotators occurs when they compare the same pair of system outputs A and B and both agree on their relative preference, i.e. (assuming that ties are allowed) either $A > B$, or $A < B$ or $A = B$. $P(A)$, therefore, equals to:

$$P(A) = \frac{|C|}{|C| + |D|} \tag{2.11}$$

where $|C|$ is the number of concordant comparisons and $|D|$ is the number of discordant comparisons. $P(E)$ captures the probability that two annotators would agree randomly and is computed as follows:

$$P(E) = P^2(A > B) + P^2(A < B) + P^2(A = B) \tag{2.12}$$

If we assume that the three outcomes are equally likely, then $P(E) = \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{1}{3}$. Empirical $P(E)$ calculated taking into account how often ties occur, is close to this value, e.g. $P(E) = 0.338$ for WMT15 into-English language pairs.

The standard kappa coefficient is designed to be used with categorical data. In the context of absolute scoring, the scores e.g. 1-5 from a multi-point scale are considered as different categories. Only the cases where annotators assign the same score to an MT output are considered agreement. Thus, disagreement between the score of 1 and 2 and disagreement between the scores 1 and 5 is treated in the same way. We note that a more appropriate method in this case is weighted kappa which takes into account the closeness of the agreement between categories, or the one-off version of weighted kappa, which ignores dis-agreements unless they are larger than one category (Artstein

and Poesio, 2008). Even though Kappa coefficient can be applied to discrete categories, 1-5 adequacy evaluation is interpreted as a gradual scale by human judges.

In either evaluation setting, the levels of inter-annotator agreement reported in the major evaluation campaigns tend to be low. The consistency of human annotation is a serious issue in MT evaluation. Due to the differences in backgrounds and experience with MT, individual annotators have different expectations and preferences and assign different scores to the same MT output. Evidently, inconsistency in manual evaluation constitutes a problem not only for the MT development, but also for the design of automatic evaluation metrics. The performance of evaluation metrics is assessed by measuring the correlation with human judgments. If the latter are not reliable, no meaningful conclusions regarding the performance of the metrics can be made.

A solution to this problem is to allow the scores coming from different annotators to contribute to an overall estimate of the quality of a given translation. Denkowski and Lavie (2010c) discuss different methods to get a single score from multiple crowd-sourced human assessments for the same MT output. Graham et al. (2013) developed a methodology for quality control of the crowd-sourced multiple assessments and suggest using their average as golden truth.

2.3 Error Analysis

One of the limitations of reference-based evaluation metrics is the lack of interpretability. As noted by Federico et al. (2014), reference-based evaluation lacks an explicit modeling of the notion of error, reducing quality to the similarity with the reference. In this regard, a research direction gaining increasing interest is error analysis and diagnostic MT evaluation (Toral et al., 2012). Various MT error typologies have already been developed for manual annotation (Vilar et al., 2006; Farrús et al., 2010; Lommel et al., 2014).

Recent studies have analyzed the impact of different translation errors either on human perception of general translation quality (Federico et al., 2014; Lommel et al., 2014; Popovic et al., 2014) or on post-editing effort (Temnikova, 2010; Blain et al., 2011; O'Brien, 2011). For example, Federico et al. (2014) propose a statistical framework based on mixed-effect models for analyzing the impact of different error types on human perception of quality and on MT evaluation metrics results. The study reveals that this impact is language-dependent and is not directly related to the frequency of translation errors, as has been previously assumed.

We note that annotation and analysis of MT errors can help establish priorities in system development and also potentially provide useful information for the traditional reference-based MT evaluation, since different types of errors have different impact on the overall MT quality as perceived by human judges and an appropriate detection and parametrization of translation errors can result in a more accurate automatic evaluation.

2.4 Summary

In this Chapter we presented the major approach to automatic MT evaluation, which is based on the assumption that a good MT must be similar to a human translation. This idea is very useful in practical terms, since it allows for rapid and cost-effective evaluation based on a straightforward comparison of the words found in the candidate and reference translations. However, this comes at a cost of certain limitations. The first is related to the fact that the metrics make simplifications that are very similar to the ones made by statistical MT systems, and therefore are unable to properly detect the errors that arise from these simplifying assumptions. Secondly, there are many possible ways of rendering the contents of the source text in the target language, which makes the evaluation based on one or even various references a challenging task.

We described the work in automatic evaluation dedicated to the improvement of evaluation accuracy, focusing on three major lines of work: the integration of linguistic information, the use of alternative resources and the employment of machine learning techniques in order to combine different metrics or different measurements.

Finally, we discussed the existing methods of meta-evaluation, as the upper limit for the performance of the metrics. The consistency of human judgments needs to be improved in order to assess further advancements in metric development. A sound approach in this sense is to use the available data but to be aware of its limitations and to work on the improvement of its reliability.

Chapter 3

ASSESSING THE IMPACT OF TRANSLATION VARIATION ON AUTOMATIC MT EVALUATION

As we have seen in the previous Chapter, guided by practical and technical considerations, MT evaluation typically relies on a human translation instead of the source text as a benchmark for comparison. A well known issue arising from this approach is the failure of automatic evaluation metrics to recognize acceptable differences between MT output and the reference, resulting in unjustifiably low scores for acceptable translations that happen to be different from the reference provided. The fact that there is no single correct solution to the translation task is widely accepted in the field, and a variety of methods have been proposed to alleviate this issue, as discussed in the previous Chapter (Section 2.1.2).

However, in our view, a related issue that makes reference-based evaluation less reliable has been largely disregarded in the literature. As we will see in this Chapter, a distinctive property of human translation is that the changes introduced to the original text go far beyond the ones dictated by the typological differences between languages. As a result, at the level of individual sentences, human translation can be different from the source in any linguistic aspect and still be considered perfectly acceptable. Different translation variants, therefore, may vary in terms of their relation to the original text and, in our view, their appropriateness for the purposes of MT evaluation.

In this Chapter we first present the concepts developed in the field of translation studies to describe and explain the differences between original and translation. In particular, we focus on the notion of translation shifts defined as optional changes introduced in translation with respect to the source text.

Next, we describe a small-scale experiment meant as a proof of concept where we empirically showed the impact of the characteristics of human translation on the results of automatic evaluation. To that end, we designed a prototype paraphrase generation

system based on a set of syntactic transformation rules that model structural changes contained in the reference translation in order to generate close translation alternatives to be used as additional references for MT evaluation.

3.1 Main Concepts

It is beyond the scope of this thesis to give a thorough presentation of the field of translation studies. Below we briefly present some important concepts from translation theory that in our view can contribute to the improvement of the practice of MT evaluation.

In the first place, Section 3.1.1 presents the concept of translation equivalence. Translation has to re-express the message of the original using the resources of the target language which often conventionalizes conceptual categories in ways that are different from the source language. This "non-isomorphism" of the language units leads to the inevitable gains and losses in translation (Szymańska, 2011). The contents of the source text cannot be fully expressed in the target language, which gives rise to a multiplicity of translation choices regarding the ways in which the equivalence between the original and translated texts can be achieved.

Depending on the priorities established by the translator, different aspects of the original text may be changed or preserved. More often than not, translators strive for functional equivalence, which implies changing the make-up of the source text far beyond strict necessity. The changes or shifts in translation that are not dictated by grammaticality are a key notion in describing the relation between the original and translated texts and has received much attention in translation studies. We turn our attention to this phenomenon in Section 3.1.3.

Translation shifts are conditioned by a wide range of linguistic and extra-linguistic factors. It has been claimed that translation as a complex cognitive process that involves interpretation and re-expression of the source text in the target language systematically affects translated texts independently of the language pair. Section 3.1.2 provides a brief discussion of the distinctive features of translated texts as compared to the texts originally written in the target language.

3.1.1 Translation Equivalence

Equivalence is one of the key concepts in translation studies that describes the relation between the original and translated texts. Various types of equivalence have been suggested based on the various aspects in which source and target texts can be equivalent. Thus, the well-known classification developed by Koller (1989) includes denotative or referential equivalence (invariance of content, the source and target words refer to the same thing in the real world), connotative or stylistic equivalence (involving a range of subtle factors such as register, sociolect, dialect, emotional marking etc., the source and

target words are assumed to trigger the same associations in the minds of source and target audience), text-normative equivalence (involving the norms of language use in a particular text type/genre), pragmatic or dynamic equivalence (source and target texts having the same effect on their respective readers) and formal equivalence (related to the form and aesthetics of the text, source and target words having similar orthographic or phonological features). In similar terms, Baker (1992) considers equivalence at various levels of linguistic description, from word level, through grammar, thematic and information structure to text cohesion and pragmatics, i.e. interpretation of text in context. Nida (1964/2000) introduced the well-known distinction between formal equivalence, which implies revealing as much as possible about the content and the form of the source text, and dynamic equivalence, i.e. the equivalence of the expected effect on the recipient. The different levels of correspondence between the original and translated texts are often interpreted as a hierarchy of priorities, with pragmatic or functional aspect occupying the highest place in the hierarchy (Szymańska, 2011). In other words, translators readily sacrifice formal correspondence between source and target language units for the sake of pragmatic equivalence.

Another important distinction concerning the types of equivalence in translation relates to the unit of comparison. Translation equivalence can be local or global, i.e. it can be established at the level of words and sentences, but also at the level of text as a whole. In human translation, local equivalence is often given up for the sake of the naturalness of the translated text in the target language. For instance, part of one sentence can be expressed in a different sentence. This type of source-target correspondence cannot be expected from MT that currently operates at sentence level only, although some work has already been accomplished in introducing discourse information into MT (Hardmeier, 2014).

3.1.2 Translationese and Translation Universals

Besides linguistic and situational factors that affect translator's behavior, it has been claimed that some properties of the translation process have a systematic impact on translators' decisions independently of the language pair. Universals of translation are linguistic features that typically occur in translated rather than in original texts and are considered to be independent of the influence of the specific language pairs involved in the process of translation (Baker, 1993, p.243). Some of the postulated universals are:

- Simplification – a tendency to make the text lexically and syntactically simpler.
- Explicitation – a tendency to spell things out rather than leave them implicit in translation. This stems on the one hand, from the target audience lacking the background knowledge that may be necessary to interpret the text, and on the other hand, from the nature of translation process itself, that involves interpretation and re-expression of the source text.

- Normalization – a tendency to conform to patterns and practices which are typical of the target language, even to the point of exaggerating them.

- Interference – transfer to the target text of the linguistic make-up of the source.

The explicitation hypothesis is especially interesting, as an illustration of the problem of reference-based MT evaluation. It has been suggested that the process of interpretation of the source text by the translator leaves traces in the target text. In other words, translators express explicitly the information that they inferred from the original text. This cannot be expected from MT systems, unless such explicitation occurs regularly given certain textual context. As a result, if the only information for the assessment of MT quality is its similarity to a reference translation, the differences between candidate and reference that result from explicitation in human translation will be treated as omission errors of the MT.

The influence of these properties of translated texts on statistical MT training has been extensively discussed in the literature (Kurokawa et al., 2009; Lembersky et al., 2012). This issue, however, has barely been explored in the context of MT evaluation.

3.1.3 Translation Shifts

The concept of translation shifts is one of the central ideas in translation studies that has received much attention throughout the history of the discipline (Catford, 1965; van Leuven-Zwart, 1989; Vinay and Darbelnet, 1958/1995; Toury, 2004), as translation theories are ultimately interested in what makes a target text "depart" from the original in certain aspects and to a varying degree. It is beyond the scope of this thesis to provide an exhaustive description of the existing theoretical approaches to this phenomenon. See (Cyrus, 2009) for a thorough overview.

The term "translation shift" was first introduced by Catford (1965). Catford's definition of this concept relies on his distinction between formal correspondence and textual equivalence. Catford (1965) defines formal correspondence as an abstract relationship that holds between two linguistic categories that occupy approximately the same place in their respective language systems, whereas textual equivalence is a relationship between source and target texts (or their parts) in a particular context. The basis for textual equivalence is the interchangeability of linguistic elements in a given situation. Thus, the distinction between formal correspondence and textual equivalence mirrors the difference between language systems and language use. Translation shifts are departures from formal correspondence between source and target language units for the sake of textual equivalence. They tend to occur when formally similar source and target constructions have different semantic and / or pragmatic values. Translation shifts are conditioned by a multitude of factors, some of which are linguistic and others extra-linguistic in

nature: typological[1], usage-based (e.g. the difference in the frequency of use of certain expressions given a text type or genre in source and target languages), socio-cultural (e.g. difference in the background knowledge of the source and target audiences) and cognitive (e.g. explicitation hypothesis from the previous Section) (Steiner, 2002).

Various typologies of translation shifts have been designed. As an example, consider the following classification by Cyrus (2006) based on a well known work by van Leuven-Zwart (1989). The annotation is based on predicate-argument structures. Each predicate and each argument represent a unit of comparison.

- Grammatical Shifts
  - **Category Change**. Translation units belong to different syntactic categories (e.g. a verbal predicate translated as a nominal predicate, i.e. nominalization)
  - **(De)passivization**. An active predicate from the source sentence has been rendered as a passive predicate in translation, or vice versa.
  - **(De)pronominalization.** The source argument is realized by lexical material (or a proper name) but translated as a pronoun.
  - **Number Change**. The corresponding translation units differ in number, i.e. one is singular and the other plural.
- Semantic Shifts
  - **Semantic Modification**. Source and target language units are not straightforward equivalents of each other because of some type of semantic divergence, for example, a difference in aktionsart between two verbal predicates.
  - **Explicitation**. The target language unit is lexically more specific than the source counterpart. This may happen either if extra linguistic material has been added in translation or if a word with a more specific meaning has been used. For example, when translating a noun, an adjectival modifier can be added in translation, or a noun standing in a hyponym relation with the corresponding source word may be used.
  - **Generalisation**. Also referred to as implicitation (Becher, 2011). Opposite of the above, i.e. the target language unit is lexically less specific than its source or some information has been left out in the translation.
  - **Addition/Deletion**. As a consequence of the annotation scheme that relies on predicate-argument structure, these categories are added in order to account for cases when arguments or predicates are added or omitted in translation.

[1]Besides the obvious differences between languages, more subtle distinctions concerning word order preferences, information structure, information density, etc. can induce structural shifts in translation (Doherty, 2006, p.13).

- **Mutation**. Source and target units are textually equivalent but they differ radically in their lexical meaning. As an example, consider the relation between two sentences that are semantically different but pragmatically equivalent: "It is cold in the room" vs. "Please, close the window".

Translation shifts are not a rare phenomenon in human translation. They are viewed by translation scholars as "parts and parcel of high quality translation" (Ahrenberg and Merkel, 2000, p.45) or even "a defining feature of translation" (Toury, 2004, p.22). Professional translators are *expected* to change the source text in order to adapt it to the norms and regularities of the target language use depending on the text type, genre, register, means of communication, etc. Even in typologically related languages, where formally similar constructions are available in many cases, they are used in different contexts in accordance with language-specific linguistic regularities.

The relation between original and translated texts has been neglected in MT evaluation. As a notable exception, Ahrenberg and Merkel (2000) developed correspondence measures to describe structural and semantic distance between source and target texts. The measures are based on the number of shifts in translated texts. Shifts were annotated manually following a classification proposed by the authors. Similarly to the shift typology designed by Cyrus (2009), Ahrenberg and Merkel (2000) make a distinction between semantic shifts and structural shifts. Regarding the semantic aspect, they distinguish between the use of more specific expression, less specific expression or expression with different meaning. Whereas Cyrus (2009) focuses on a very specific types of structural changes, Ahrenberg and Merkel (2000) present a more detailed classification covering a wide number of syntactic changes, which we reproduce below.

- Changes related to the function and properties of clauses:
  - Voice shift (e.g. active $>$ passive)
  - Sentence mood shift (e.g. imperative $>$ declarative)
  - Shift of finiteness (e.g. finitival > infinitival verb construction)
  - Level shift (e.g. main clause $>$ subordinate clause, clause $>$ phrase)
  - Function shift (e.g. temporal clause $>$ conditional clause)
- Changes related to the function and position of constituents:
  - Function shifts (e.g. manner adverbial $>$ predicative)
  - Level shifts (e.g. phrase $>$ clause)
  - Transpositions (changes in order between constituents)
- Changes related to the number of syntactic constituents:
  - Additions
  - Deletions
  - Divergences (1 source constituent $>$ 2 or more target constituents)
  - Convergences (2 or more source constituents $>$ 1 target constituent)

- Paraphrases, which influence at least two constituents and cannot be split up into several smaller changes

Based on translation shifts annotation conducted using the above typology, Ahrenberg and Merkel (2000) computed correspondence measures for different text genres and different types of translation: human translation, computer assisted human translation and MT. The results of the study confirmed that MT is formally closer to the source text than human translation. This is not surprising given that (a) in statistical MT alignments involving short phrases are more frequent and therefore possible translation options are limited by low-level correspondences; and (b) translator's decisions are shaped by extra-linguistic factors inaccessible to the MT systems.

The next Section presents our experiment that investigates the impact of translation shifts on MT evaluation (Fomicheva et al., 2015a). Instead of comparing the number of shifts in MT and human translation, we automatically paraphrase the latter to produce close translation variants and test if using them along with the reference translation results in higher automatic evaluation scores and more accurate automatic evaluation.

3.2 Impact of Translation Shifts on Automatic Machine Translation Evaluation

In our preliminary experiment we analyzed the impact of optional translation shifts present in the reference translation on automatic MT evaluation. To that end, we developed a prototype paraphrase generation system based on a set of hand-crafted transformation rules that "undo" optional shifts in the reference translations. The generated paraphrases are then used for automatic evaluation as additional references and the results of single-reference and multi-reference evaluation are compared.

We do not use any information from the source sentence, as the generated references are to be used together with the original human reference. If human translator has changed a source construction that is preserved in MT, using the relevant paraphrase increases automatic evaluation score. If this is not the case, the additional reference is not supposed to have any effect on the evaluation.

Translation shifts are language-specific. For this experiment, we worked with English–Spanish translation. We focused on the syntactic aspect of translation, since lexical variation has been already extensively addressed in automatic MT evaluation (see Section 2.1.2).

3.2.1 Translation Shifts Typology

As we aim to generate translation alternatives in the target language, we studied available formal descriptions of linguistic paraphrase (Barrón-Cedeño et al., 2013; Bhagat

and Hovy, 2013) as well as translation shifts classifications (van Leuven-Zwart, 1989; Cyrus, 2006; Ahrenberg, 2005) as the basis for the design of transformation rules for the generation of close translation variants.[2] Based on these works, we developed the following typology of translation phenomena, which was used for the design of transformation rules and for the evaluation of our paraphrase generation system.

1. Grammatical features (e.g. finiteness, mood, modality, tense, aspect)
2. Grammatical category (e.g. pronominalization, nominalization, manner adverbial $\rightarrow$ predicative adjective)
3. Diathesis (e.g. passive construction $\rightarrow$ active construction, personal clause $\rightarrow$ impersonal clause)
4. Level/function (e.g. phrase $\rightarrow$ clause, main clause $\rightarrow$ subordinate clause, temporal clause $\rightarrow$ conditional clause, locative-possessive alternation)
5. Word order (e.g. subject-predicate inversion, clitic climbing, changes in the position of adverbial modifiers)
6. Number of constituents (e.g. ellipsis, explicitation/implicitation)

Individual changes that are entailed by other operations are not considered separately. For example, the inversion of arguments induced by a diathesis alternation is not annotated as a change in word order.

3.2.2 Reference Generation

The process of paraphrase generation is illustrated in Figure 3.1. The system operates on dependency trees in CONLL format (Buchholz and Marsi, 2006) and returns full transformed sentences.[3] In the pre-processing stage, dependency parses of the reference translations (HRTs in Figure 3.1) are produced. In the analysis stage, the structures to be transformed are identified by means of regular expression matching. In addition to syntactic information, a grammatical dictionary of Spanish Resource Grammar (Marimon, 2013) with verb frame information is used to introduce lexical restrictions for rule

[2]There is a close relation between translation and paraphrasing. In principle, there are many paraphrases of the original sentence and the question for the translator is: which of them would recreate the author's choice in another language. The experiments described in this Section were inspired by the so called "control paraphrase" method for the analysis of translation shifts proposed by Doherty (2006), which consists in gradually, through a set of paraphrases, generating the actual translation from an "analogous" translation that reproduces the original sentence as closely as possible.

[3]In our experiments, we used the MaltParser dependency parser (Nivre et al., 2007) with Spanish models (Marimon et al., 2014) available at `http://www.iula.upf.edu/recurs01_mpars_uk.htm`.

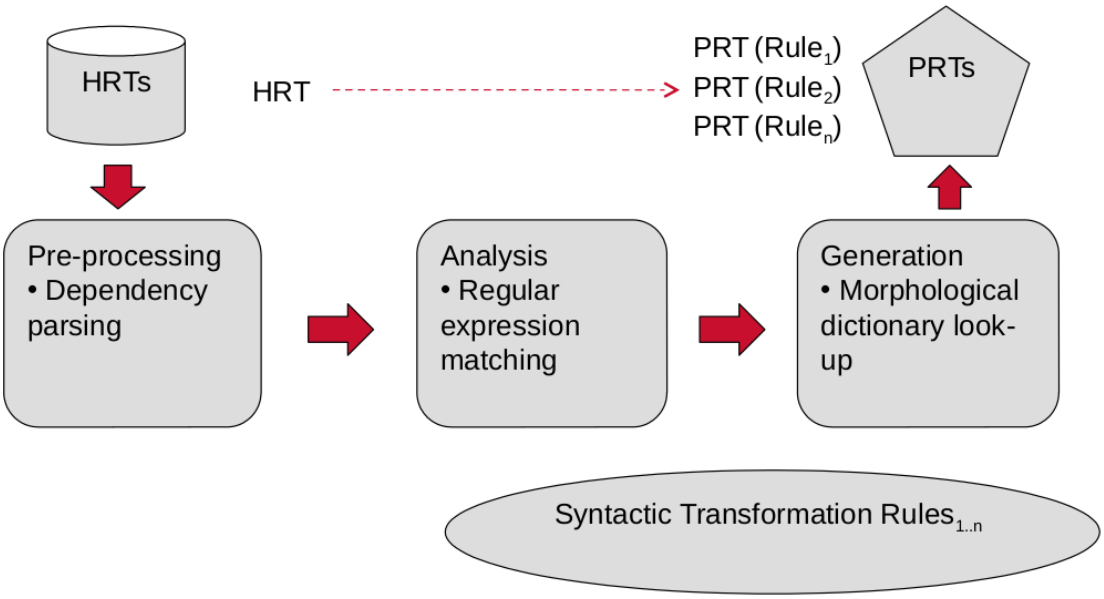


Figure 3.1: Generation of Paraphrased Reference Translations (PRTs)

application. If the conditions are matched and no restrictions are found, in the generation stage the system reconstructs the sentence with relevant changes using information extracted from the parses of the input sentences and morphological dictionary look-up in order to generate the appropriate word forms. From one input sentence, the system generates a set of paraphrases (as many as rules are applied).

3.2.3 Transformation Rules

The experiment was designed as a proof of concept. Therefore, we implemented a limited set of rules that cover only a small portion of possible structural shifts in English-Spanish translation. The rules are put in relation to the structural shifts typology presented in the previous section. The selection of particular rules was motivated by a) the feasibility of the implementation and b) their relevance given the typological relations between the source and target languages (English and Spanish, respectively):

1. Grammatical Features
   (a) past simple ↔ present perfect
   (b) present simple ↔ present perfect
   (c) present simple ↔ past imperfective
   (d) simple verb form → progressive construction

(e) simple future ↔ periphrastic future

(f) recent past periphrasis → present perfect + 'recently'

(g) habitual aspect periphrasis → simple verb form + 'normally'

(h) repetitive aspect periphrasis → simple verb form + 'again'

2. Grammatical Category

   (a) nominalization ↔ denominalization

   (b) prepositional phrase → adverbial modifier

   (c) copulative clause → adverbial modifier

3. Diathesis

   (a) active → analytic passive

   (b) synthetic passive → analytic passive

   (c) personal → impersonal

4. Word Order

   (a) post-verbal subject → pre-verbal subject

   (b) post-verbal adverbial ↔ pre-verbal subject

   (c) VP-external adverbial ↔ VP-internal adverbial

   (d) sentence-initial adverbial ↔ post-verbal adverbial

   (e) sentence-initial detached PP ↔ post-verbal detached PP

   (f) pre-nominal adjectival modifier → post-nominal adjectival modifier

   (g) clitics in pre-verbal position → clitics in post-verbal position

5. Number of constituents

   (a) personal pronouns with subject function[4]

   (b) repeated preposition heads in coordinated PPs

For cases where the direction of the optional change in human translation cannot be predicted, the rules are applied in both directions (marked with bi-directional arrows above). For example, the English "-ing" verb forms with nominal function may be translated by nouns or infinitives in Spanish and in this case we cannot say that one option is closer to the source sentence than the other.

It should be noted that some of the constructions from the above list may be considered equivalent only given a specific linguistic context (for instance, tense alternations).

[4]Spanish is a pro-drop language and can elide subject pronouns. By contrast, their use is mandatory in English.

We do not use any source-side information and thus applying such rules may result in paraphrases that change the contents of the original. However, these are not supposed to affect the results because the paraphrases are to be used together with the true human reference translation. Thus, in case the human translator has changed an original construction that is preserved in MT, using the relevant paraphrase increases the automatic evaluation score. If it is not the case, the additional reference is not supposed to have any effect on the evaluation.

Many of the relevant operations described in Section 3.1.3 cannot be modeled computationally for the same reasons that we cannot expect to find these changes in the MT output. For example, in cases of implicitation shifts where content is left unexpressed, we lack information to reconstruct it. Furthermore, no high quality language processing tools are yet available for analyzing certain complex phenomena. For example, in case of pronominalization shifts, when a full noun phrase is substituted by a pronoun, which frequently happen in translation in order to avoid repetition, we lack high quality co-reference resolution tools to reconstruct the original full noun phrase.

3.2.4 Dataset

For the experiments, we used the EAMT09 dataset (Specia et al., 2010a) that consists of 4,000 source sentences in English randomly extracted from Europarl (Koehn, 2005), their corresponding human translations into Spanish and MT outputs produced by 4 statistical MT systems (see Section 2.2.2). The Europarl corpus is a multilingual, parallel corpus that has been collected from the proceedings of the European Parliament since 1996 and includes translations from (into) 21 official languages of the European Union. Translation is done by professional translators, who are native speakers of the corresponding target languages (van Halteren, 2008).

The EAMT09 dataset (Specia et al., 2010a) includes manual evaluation scores, collected from professional translators using the post-editing criterion. Specifically, the annotators were asked to indicate the amount of editing needed in order to make the MT ready for publishing: 1) requires complete re-translation, 2) a lot of post-editing is needed, 3) little post-editing is needed, 4) fit for purpose. It is important to note that human assessments were obtained independently from the reference translation, from bilingual speakers, using the source text and the MT output. Therefore, the presence of translation shifts in the reference translation could not have any effect on the results of this manual evaluation.[5]

We must note that Europarl proceedings are published in all of the official languages of the European Union, allowing to create a multilingual and parallel corpora for the language pairs involved. In the parallel datasets of the Europarl corpus the actual translation direction is ignored, as they include sentences that were translated into the target

[5]The problem of reference bias in manual evaluation is discussed in detail in Chapter 6.

language from any source language (Koehn, 2005). For example, the English-Spanish dataset contains translations into Spanish (and into English) from any other language of Europarl proceedings, or translations in the opposite direction, i.e. from Spanish into English.

In the experiments described in this Chapter, we were interested in modeling prototypical structural shifts in English–Spanish translation direction. Translation preferences for structural shifts are necessarily language-specific. We, therefore, used the original Europarl corpus where the original language of the speaker is indicated[6], in order to verify how many of the sentences of the EAMT09 dataset are actually translations from English into Spanish. We found out that out of 4,000 sentences, 2,976 (74%) are actual English–Spanish translations. We consider that this is a reasonable amount, although the fact that part of the data comes from different translation directions may have affected the results reported in Section 3.2.5.

We must note that, until recently, it has been common practice in the evaluation campaigns to ignore translation direction. The test sets for MT development created based on Europarl (Koehn, 2005; Koehn and Monz, 2006) for various language pairs included sentences translated to the target language from any source language. After substantial work on the influence of translationese on MT training (Kurokawa et al., 2009; Lembersky et al., 2012), this practice has been deprecated (Bojar et al., 2014).

3.2.5 Experimental Results

The paraphrase generation system was applied to the reference translations. A separate reference translation was generated for each of the groups of rules presented in Section 3.2.3. MT outputs were automatically evaluated with BLEU in a single reference baseline scenario and in a multi-reference scenario, with automatically generated paraphrases. When multiple references are provided, BLEU takes into consideration n-gram matches between the candidate translation and each of the references. Thus, in principle, the impact of different groups of transformation rules (corresponding to different types of translation shifts discussed in Section 3.2.1) on MT evaluation can be assessed. Note that the aim of these experiments was not to improve the accuracy of automatic MT evaluation (broad coverage paraphrase generation tools discussed in Section 2.1.2 are more suitable for this purpose), but to test the effect of specific translation phenomena discussed in the theoretical part of this Chapter on reference-based automatic evaluation.

[6]The information regarding the original language of the speaker is not always available (Halteren van, 2008; Islam and Mehler, 2012). We counted only those sentences that were tagged in Europarl as translated from English.

	Reference	Morphology	Structure	Order	Number	All
System 1	.354	**.357**	**.356**	.355	.355	**.359**
System 2	.326	.329	.325	**.329**	.327	**.332**
System 3	.355	.355	.355	.356	.353	.356
System 4	.247	**.251**	.247	.247	.247	.251

Table 3.1: Sentence-level Pearson correlation for BLEU evaluation in single-reference and in multi-reference scenarios

	Morphology	Structure	Order	Number	All
System 1	23.7 %	12.3 %	11.5 %	11.1 %	39.1 %
System 2	22.8 %	13.6 %	10.7 %	12.6 %	39.3 %
System 3	26.6 %	13.5 %	11.1 %	14.5 %	44.1 %
System 4	27.0 %	14.4 %	10.8 %	13.6 %	43.5 %

Table 3.2: Percentage of affected sentences for BLEU evaluation in single-reference and in multi-reference scenarios

Extrinsic Evaluation

We compared the Pearson correlation with human judgments of single-reference BLEU and multi-reference BLEU with our additional references. Table 3.1 shows Pearson correlation with human judgments using the original human reference (**Reference**) and the original reference together with the additional references generated using the different groups of transformations rules separately (**Morphology**, **Structure**, **Order** and **Number**), as well as in combination (**All**). Here and in the rest of this work we use the Hotteling-Williams test for testing statistical significance in dependent correlations (see Section 2.2.3). Results found to be significantly different from the single-reference correlation are marked in bold.[7]

The results show that there is a significant improvement in correlation when using all the rules for systems 1 and 2, but not for systems 3 and 4. Note that, as shown in Table 3.2, the percentage of sentences with BLEU scores affected by the application of the rules is quite low. This is not surprising, since we modeled a very limited number of translation phenomena. The fact that the differences in correlation is still significant means that the changes introduced by the rules are appropriate in the majority of the cases. As far as the difference between the MT systems is concerned, as shown in Table 3.3, the average quality of the translations produced by the systems 1 and 2 is higher,

[7]The correlation results for System 4 are the same for the groups "Morphology" and "All". However, only for the former the difference with the same-reference scenario is significant. Recall from Section 2.2.3 that besides the correlation with human judgments, the Hotteling-Williams test takes into consideration the correlation between the two metrics that are being compared.

which explains the results. The scores may be increased when using multiple reference due to noisy matches between the words that are actually unrelated in the two sentences. Since the number of low quality translations is higher in the case of systems 3 and 4, the noise introduced by the multiple references affects the correlation more strongly.

	Human	**BLEU**
System 1	2.83	0.4018
System 2	2.56	0.3641
System 3	2.51	0.3351
System 4	1.34	0.2018

Table 3.3: Average human scores and system-level BLEU scores for the MT systems 1-4 from the EAMT09 dataset

Intrinsic Evaluation

We performed a detailed analysis of the impact of different types of translation shifts on MT evaluation. To that end, we performed manual annotation of translation shifts on a sample of the data. We randomly selected 290 sentences from the data.[8] Optional structural shifts in reference translations were annotated and classified manually using the typology presented in Section 3.1.3.

Single-reference and multi-reference BLEU was again computed on this sample of the data. We then used the BLEU scores to calculate precision and recall for the transformation rules as follows. The purpose of using additional references was to increase BLEU scores for cases when a translation shift occurs in human translation while MT contains the corresponding structure that is formally equivalent to the source. Therefore, for each group of rules we considered that the application was successful if using the respective set of paraphrases increases the BLEU score and the corresponding translation shift occurs in human translation (true positives).

By contrast, rule application was considered unsuccessful when no translation shift of a certain type occurs in human translation and applying the corresponding set of rules increases the evaluation score (false positives), or when there is an optional change in the reference and applying the corresponding set of rules does not increase BLEU score (false negatives).

In order to assess the performance of the system per se, we counted recall separately for all the annotated translation shifts vs. translation shifts modeled by the rules. The results are shown in Table 3.4. The overall precision indicates that in 70% of cases

[8] We made sure that the sentences were originally translated from English into Spanish (see discussion in Section 3.2.4).

Rules	P	R_{shifts}	R_{rules}	Freq
Grammatical features	0.76	0.43	0.60	104
Grammatical category	0.70	0.30	0.61	77
Diathesis	0.59	0.20	0.43	66
Word order	0.79	0.40	0.72	151
Number of constituents	0.61	0.23	0.56	82
Total	0.69	0.32	0.58	480

Table 3.4: Precision (P) and Recall (R) for the application of transformation rules and Frequency (Freq) of translation shifts

of rule application the system successfully reconstructs a close translation option and using it as additional reference increases the BLEU score. As expected, recall is low in case all translation phenomena are considered, and much higher if calculated only for the phenomena covered by the rules. Thus, the system shows good performance in the cases it is designed to deal with. The overall number of translation shifts is high as there is an average of 1.7 optional changes per sentence in the reference, confirming the idea that such changes are indeed common practice in human translation.

An example of a successful rule application is given in Table 3.5. In this example a clause-level adverbial modifier is changed into predicative adjective in human translation (with corresponding changes in sentence structure). This transformation is common in English–Spanish translation, as translators are advised to avoid excessive use of manner adverbials in -mente (-ly) considered a calque from English where they are more frequent. MT preserves the structural organization of the original, which results in a sentence that may be considered stylistically flawed, but is perfectly acceptable according to the human evaluation score. The paraphrase generated by the system successfully neutralizes this shift in human translation, and using it increases the BLEU score. Note, however, that the increase is small as the exact position of the adverbial is still not matched.

As far as specific groups of rules are concerned, the lowest results are for diathesis changes. In this group the most frequent transformation is reflexive passive $\rightarrow$ analytic passive. The resulting paraphrases are irrelevant, as they do not increase the BLEU score because the corresponding shift frequently occurs in the MT. This is understandable given the nature of statistical MT. Since the change only involves local context and is frequently present in English–Spanish translations, it can be expected to occur in MT. By contrast, word order changes are more challenging for statistical systems. For this reason, the group of rules that neutralize the optional changes affecting word order obtained the highest precision and recall.

Source:	this event, on the eve of the lahti meeting, is clearly of particularly crucial significance to us.
MT:	este acontecimiento, en vísperas de la reunión lahti, es claramente de especialmente crucial importancia para nosotros. *this event, on the eve of the lahti meeting, is clearly of particularly crucial significance for us.*
Ref:	está claro que este acontecimiento, en vísperas del encuentro de lahti, reviste para nosotros una especial trascendencia. *it is clear that this event, on the eve of the lahti meeting, represents for us a special importance.*
PRT	claramente, este acontecimiento, en vísperas del encuentro de lahti, reviste para nosotros una especial trascendencia. *clearly, this event, on the eve of the lahti meeting, represents for us a special importance.*
	BLEU
Ref	0.2477
PRT	0.2670

Table 3.5: Example of category change in human translation (EAMT09 dataset, English–Spanish translation, sentence 1007)

As an illustration, consider the example shown in Table 3.6.[9] Here the reference contains two optional changes: the transformation from analytic passive to reflexive passive and subject-predicate inversion. The first paraphrase in Table 3.6 (PRT_1) delivers the close version with analytic passive construction. The second paraphrase (PRT_2) reconstructs the word order of the source sentence neutralizing the subject-predicate inversion present in human translation. In the case of word order, the rule is applied successfully as it increases the BLEU score. In the case of diathesis transformation, the shift occurs in both human translation and MT and thus the transformation performed by our paraphrase system is not relevant.

Another source of errors is that, contrary to our assumption, not using source-side information does introduce noise. This is the case, for example, when the transformation involves adding a function word that happens to be present in MT but does not form part of the same syntactic construction.

Finally, both precision and recall are affected by parser errors. For instance, word order changes cannot be addressed in cases where the parser fails to identify the head of the element to be moved. Parser errors are especially harmful for rule-based approach as the patterns have to be defined in detail and the conditions need to be exactly satisfied for the rules to apply.

[9]MPASS stands for the Spanish passive marker "se".

Source:	appropriate arrangements have been made for consultation with the member states.
MT:	los preparativos apropiados se han hecho para su consulta con los estados miembros. *appropriate arrangements MPASS have made for the consultation to the member states.*
Ref:	se han realizado los preparativos apropiados para la consulta a los estados miembros. *MPASS have made appropriate arrangements for the consultation to the member states.*
PRT_1	han sido realizados los preparativos apropiados para la consulta a los estados miembros. *have been made appropriate arrangements for the consultation to the member states.*
PRT_2	los preparativos apropiados se han realizado para la consulta a los estados miembros. *appropriate arrangements MPASS have made for the consultation to the member states*
	BLEU
Ref	0.3013
PRT_1	0.3013
PRT_2	0.4683

Table 3.6: Example of diathesis change and subject-predicate inversion in human translation (EAMT09 dataset, English–Spanish translation, sentence 1283)

3.3 MT Evaluation via Post-Editing

As an interesting piece of evidence regarding the impact of the variation in human translation of the performance on automatic evaluation metrics, we compared the results of automatic evaluation when human translation vs. the post-edition of MT are used as references.

We used the EAMT11 dataset containing 2,525 French sentences in the news domain and their translation into English produced by a baseline phrase-based MT system (Specia, 2011), as well as a human reference translation. The data was randomly extracted from the news-test2009 dataset (Callison-Burch et al., 2010). The EAMT11 dataset also contains the post-edition of the MT output provided by professional translators, as well as human scores reflecting post-editing effort on a scale from 1 to 4. Table 3.7 shows the average scores produced by the commonly used evaluation metrics when using human translation and post-edition for automatic evaluation. The scores computed with

the post-edition are higher by a very large margin, which indicates that a substantial portion of the differences between candidate and reference translations are not due to translation errors. Note that this is the case not only for BLEU, which only allows for exact word matching, but also for Meteor, that recognizes synonyms and paraphrases. Therefore, current mechanisms for addressing acceptable variation clearly do not suffice for reference-based MT evaluation, which we relate to the presence of translation shifts in human translation.

	Translation	**Post-Edition**
BLEU	0.2368	0.6579
Meteor	0.3259	0.5659
TER	0.6294	0.2251

Table 3.7: Average scores of BLEU, Meteor and TER metrics for the MT outputs from the EAMT11 dataset using human translation (Translation) and post-edited MT (Post-Edition) as reference

The EAMT11 dataset presents another interesting illustration of the effect of using reference translation on MT evaluation. Human translations extracted from the newstest2009 dataset (Callison-Burch et al., 2010) actually come from different original languages, whereas MT outputs are French-into-English translations. Table 3.8 reports average metric scores (Avg) and Pearson correlation (r) for the sentences coming from different original languages in the EAMT11 dataset, the last column shows the number of sentences. Clearly, both the average score and Pearson correlation with human judgments is substantially higher when both the MT output and human translation were generated from the same source language. It has been shown that taking into account translation direction is highly beneficial for statistical MT training. However, this has not been fully understood in the context of MT evaluation. In this work, we argue that a careful consideration of the properties of human translation used as reference is needed in order to achieve more accurate MT evaluation.

3.4 Summary

In this Chapter, first, we have provided a discussion of translation variation from the point of view of translation studies. We established that in human translation a common practice is to introduce optional changes with respect to the original text. Human translation tends to shift away from the original sentence paraphrasing what could be considered as a close translation variant. By contrast, MT tends to be close to the original, as translation choice is limited to frequent correspondences and is conditioned by local context. Thus, in part, linguistic variation between different translation options results from the presence of translation shifts. In human translation the shifts occur due to

	BLEU		Meteor		TER		
	Avg	r	Avg	r	Avg	r	Sentences
it	0.264	0.2273	0.3417	0.2989	0.5433	-0.2056	299
en	0.2299	0.2583	0.3329	0.2889	0.6349	-0.19	370
hu	0.1583	0.3385	0.287	0.3783	0.7656	-0.2632	367
es	0.1874	0.2602	0.2774	0.1519	0.6847	-0.1923	345
fr	**0.431**	**0.3734**	**0.4463**	**0.4312**	**0.3849**	**-0.3474**	352
cz	0.2288	0.2554	0.3215	0.2386	0.6387	-0.1906	434
de	0.1678	0.2969	0.2791	0.3346	0.7321	-0.2806	358

Table 3.8: Average scores and Pearson correlation with human judgments for sets of sentences from the EAMT11 dataset that come from different sources languages

a wide range of linguistic and extra-linguistic factors. Some of them are in fact beyond the reach of current MT systems. As a result, in the context of MT evaluation, the differences between candidate and reference translations can be related to the idiosyncrasies of human translation, in which case different translation choices found in the MT and the reference are considered equally valid in manual quality assessment (however, see discussion in Chapter 6), but tend to be harshly penalized by automatic evaluation metrics. While it can be argued that MT ultimately aims at delivering the same results as professional translation, a proper understanding of the nature of the differences between MT and human translation is needed in order to establish priorities in MT development.

Second, we conducted a study meant as a proof of concept to illustrate the impact of translation shifts in the reference translation of the results on automatic MT evaluation. We devised a prototype paraphrase generation system based on a set of syntactic transformation rules that models structural changes contained in the reference translation in order to generate close translation alternatives to be used as additional references for MT evaluation. We showed that the presence of translation shifts in the reference translation has a negative impact on the accuracy on automatic MT evaluation, as reflected by human judgments. The experiments suggest that using a close reference translation may result in better MT evaluation.

We are aware, nevertheless, that the possibilities of neutralizing optional shifts in human translation producing close translation options suffers from the same limitations as automatic MT. The shifts that are challenging for MT, such as explicitation or implicitation performed based on the external knowledge, cannot be captured by the approach we discussed here for the same reasons. Some of the shifts discussed can be found in the MT provided that they are frequent and conditioned by local context.

Our experiments also revealed some practical limitations of reference generation as a method for handling the problem of acceptable variation between MT and the reference. The use of additional references results in spurious matches in cases when

generated words are parts of different constructions in the candidate and references sentences. The use of alternative references is a one-sided solution, since while it increases the coverage of possible variations accepted by the metrics, it suffers from the fact that the impact of the differences between MT and the reference(s) is not estimated properly. In the following chapters we explore alternative techniques to distinguish between acceptable variation and MT errors that result in substantial improvements in correlation with human judgments.

Chapter 4

USING WORD CONTEXT FOR AUTOMATIC MT EVALUATION

One of the defining characteristics of translation is its equivalence to the source text. Setting aside the complexity of the concept of translation equivalence (see Section 3.1.1), the assessment of the quality of human translation necessarily involves a comparison with the original text. In MT evaluation, the task is usually simplified by using a human translation as gold standard. This allows for fast and inexpensive automatic evaluation that proceeds by counting the number of matching words and word sequences between the MT output and the human reference translation. Thus, reference-based MT evaluation relies on the assumption that the more similar an MT output is to a human translation, the higher the MT quality. Candidate-reference similarities (matching words), therefore, are treated as good translation choices, whereas candidate-reference differences (missing words, added words or words placed in a different order) are assumed to be indicative of bad translation choices. Logically, a quality score can then be obtained based on the number of good and bad choices in the MT output.

Unfortunately, such a straightforward approach does not work well at the level of individual sentences. Word matches *per se* do not necessarily indicate high quality, whereas the differences are not a guarantee of MT errors. On the one hand, the differences that are considered acceptable in manual evaluation are harshly penalized by evaluation metrics.[1] On the other hand, a low quality MT output may obtain a relatively high score if it contains a high number of local lexical matches with the reference while being grammatically ill-formed or semantically abnormal. Thus, the impact of the differences between MT and a reference translation varies considerably. As discussed in the previous Chapter, they may be due to the idiosyncrasies of human translation. But they may also be indicative of incorrect translation, from minor imperfections to serious

[1] As we will see in Chapter 6, in monolingual evaluation, where human judges do not have access to the original and assess the MT quality using reference translation as a benchmark, they are also biased by the reference, but evidently to a lesser extent than the metrics and probably in different ways.

translation errors that can render the MT output unintelligible.

In the previous Chapter, we examined the sources of linguistic variation in human translation. We identified optional translation shifts as one of the sources of differences between possible translations of the same original text. We modeled prototypical translation shifts that can be found in human translation and generated close translation variants to be used as benchmark for the evaluation of MT in order to avoid penalizing candidate-reference differences related to the presence of optional translation shits in the reference. We focused, therefore, on the problem of acceptable variation obtaining a small, but significant improvement in terms of the correlation with human judgments, which evidenced the impact of the presence of translation shifts in the reference on the results of automatic evaluation.

In this Chapter, in order to achieve further improvement in the correlation with human judgments we developed a method for quantifying the impact of candidate-reference differences on translation quality in general, without limiting ourselves to the acceptable differences related to linguistic variation in human translation. The method consists in comparing local syntactic contexts of the matching candidate and reference words in order to determine to what extent the matches (or mismatches) are indicative of MT quality. Intuitively, a match between candidate and reference words is only meaningful if they play similar roles in the corresponding sentences. Therefore, we propose to weight lexical similarity between the words by the difference in their syntactic contexts. In this way, lexical matches increase the sentence-level evaluation score depending on the number of different words in their contexts. Conversely, candidate-reference differences decrease the evaluation score depending on how many matching words they affect.

We implemented this approach as a full-fledged evaluation metric, UPF-Cobalt. Inspired by the alignment-based evaluation from Meteor (Denkowski and Lavie, 2014) (see Section 2.1.1), our metric proceeds in two stages. In the first stage, an alignment between the candidate and the reference translation is established. Based on a discussion of the existing strategies for monolingual word alignment we suggest that contextually informed alignment is beneficial for MT evaluation. Specifically, we experiment with Monolingual Word Aligner (Sultan et al., 2014) that compares the syntactic contexts of the words to discriminate between possible alignments, and show that the problem of spurious matches between unrelated candidate and reference words can thus be avoided. In the second stage, the score is computed based on the nature of the matches. In this stage, besides considering the type of the match, the difference between the syntactic contexts of the matching words is computed and integrated in the overall sentence-level score. The insights from the previous Chapter regarding linguistic variation between possible translation options are integrated in the metric. Specifically, distributed word representations are used in order to increase alignment coverage for similar words thus addressing lexical shifts in human translation (see Section 3.1.3). Equivalence of some

syntactic constructions is considered through a mapping between the corresponding dependency functions when comparing the syntactic contexts of the words. We evaluated the performance of our approach using different types of human judgments. A detailed analysis highlighting advantages and limitations of the metric in various settings and configurations was conducted. Our metric outperforms most of the current metrics for automatic evaluation into English in a variety of different evaluation settings (Fomicheva and Bel, 2016; Fomicheva et al., 2016, 2015b) and beats the baseline evaluation metrics by a significant margin.

In the rest of this Chapter, we first discuss how the relation between candidate and reference words can be properly established through the use of monolingual alignment with syntactic evidence (Section 4.1). Then, we develop the idea of a syntactic context penalty to be applied to the lexical matches between candidate and reference translations (Section 4.2). Next, the UPF-Cobalt metric is presented in detail (Section 4.3). Section 4.4 contains the results of the meta-evaluation of the metric on different datasets. Finally, Section 4.5 presents a further analysis of the performance of the metric.

4.1 Monolingual Alignment

In our view, in order to compare candidate and reference sentences, it is convenient to establish correspondences between smaller units (words or phrases) first. This task is known as monolingual word alignment, where words in one sentence are mapped to semantically similar words of another sentence in the same language. We consider that such an alignment, if accurately established, can be highly beneficial for reference-based MT evaluation.

Although in automatic MT evaluation n-gram matching and alignment-based approaches are similar in practice (both intend to find how many similar words are contained in the candidate and reference translations) and the terms "matching" and "alignment" are often used indistinguishably, the alignment-based strategy has some advantages. MacCartney et al. (2006) provide an interesting discussion of the advantages of alignment over matching for comparing semantic graphs for the task of textual entailment recognition[2]. According to MacCartney et al. (2006, p.43), the simple graph matching formulation of the problem is hindered by three important issues:

> First, the above systems assume a form of upward monotonicity: if a good

[2]Textual entailment is defined informally as a relation between two natural language sentences (a premise P and a hypothesis H) that holds if a human reading P would infer that H is most likely true (Dagan et al., 2006). The textual entailment recognition task was introduced by Dagan et al. (2006). Note the similarity between this task and automatic MT evaluation. A perfect MT output and the reference translation must entail each other. This idea is exploited in the work of Padó et al. (2009) who use a textual entailment recognition system for the task of MT evaluation and obtain promising results (winning system at WMT2008 (Callison-Burch et al., 2008)).

> match is found with a part of the text, other material in the text is assumed not to affect the validity of the match [...]
>
> The second issue is the assumption of locality. Locality is needed to allow practical search, but many entailment decisions rely on global features of the alignment [...]
>
> The last issue arising in the graph matching approaches is the inherent confounding of alignment and entailment determination. The way to show that one graph element does not follow from another is to make the cost of aligning them high. However, since we are embedded in a search for the lowest cost alignment, this will just cause the system to choose an alternate alignment rather than recognizing a non-entailment.

We note that very similar problems arise in the context of reference-based MT evaluation. The differences between candidate and reference translations do not affect the contribution of the n-gram matches to the sentence-level score. Local matches (especially long ones) may result in a high overall score, even if the MT output is grammatically ill-formed. Finally, the metrics aim to find as many matches as possible, and as a result, candidate and reference words that come from different source words can end up counted as matching. If the words that resulted from translating different parts of the source sentence are aligned, then the match is meaningless for the purposes of evaluation.

Consider the example of Russian–English translation in Table 4.1. The matches (the_{2MT} - the_{4REF}), (to_{1MT} - to_{1REF}), (of_{2MT} - of_{1REF}) or (the_{4MT} - the_{2REF}) are spurious, as they resulted from translating different words in the original sentence. Spurious matches increase surface reference-based scores, although they are not indicative of semantic similarity between candidate and reference sentences or of the quality of MT output.

Source:	Лучших судьи, что называется, взяли на карандаш, чтобы к моменту завершения учебы пригласить к себе на работу.
Ref*:	*Of the best judges, so to speak, took note, so that by the moment of ending of the training invite to work.*
MT:	the_1 best judges of_1 what is called , took the_2 pencil to_1 the_3 completion of_2 the_4 study to_2 invite to_3 work .
Ref:	the_1 judges took note , so to_1 speak , of_1 the_2 very best , in order to_2 offer them jobs at the_3 end of_2 the_4 training .

Table 4.1: Example of spurious matches between candidate and reference translations (WMT2016 dataset, Russian–English translation, sentence 2897)

One way to address the problem of spurious matches proposed by Liu and Gildea (2007) is by aligning both the candidate and the reference to the source and constraining

n-gram matching to the words that are aligned to the same parts of the source sentence (see Section 2.1.3). This method, however, relies on noisy bilingual alignments. In this work, a different solution is proposed consisting in using monolingual word alignment techniques and using word context for selecting the best alignment candidates. The underlying intuition is that if the matching words occur in the same contexts, we can assume they have similar functions in the corresponding sentences and were translated from the same source words.

Using word context in order to discriminate between possible alignments has been widely explored for the task of monolingual alignment (MacCartney et al., 2006; Thadani et al., 2012; Sultan et al., 2014). However, to the best of our knowledge, this is the first time this idea is applied to the task of MT evaluation.

Syntactic context based on a dependency representation has been shown to be particularly beneficial for alignment accuracy (MacCartney et al., 2006). We selected an existing alignment tool which takes syntactic context information into account, Monolingual Word Aligner (MWA) developed by Sultan et al. (2014). Table 4.2 summarizes the results from Thadani et al. (2012) and Sultan et al. (2014) showing that MWA outperforms alternative state-of-the-art approaches on the MSR benchmark for alignment task (Brockett, 2007), including the aligner from Meteor evaluation metric.[3] Below we describe MWA in detail and provide an illustration of the advantages of contextually-informed alignment for MT evaluation.

System	F_1
Meteor	75.5
MacCartney et al. (2008)	85.3
Thadani and McKeown (2011)	87.8
Yao et al. (2013)	88.6
MWA (Sultan et al., 2014)	91.7

Table 4.2: Performance of various aligners on MSR dataset measured in terms of F1

Monolingual Word Aligner

MWA operates as a pipeline of alignment modules that differ in the types of word pairs they align: identical word sequences, named entities, content words and function words. Each module makes use of contextual evidence to make alignment decisions. In addition, the last two modules allow fuzzy matches between similar words. Named entities are aligned separately to enable the alignment of full and partial mentions (and acronyms) of the same entity. Stanford Named Entity Recognizer (Finkel et al., 2005)

[3]Meteor version from Denkowski and Lavie (2011) is referenced in the paper.

is used to identify named entities. After aligning the exact term matches, any unmatched term of a partial mention is aligned to all terms in the full mention.

MWA makes alignment decisions based on the lexical similarity and contextual evidence. The alignment score for a candidate word pair is calculated as a weighted sum of lexical and contextual similarity. The lexical similarity component identifies word pairs that are possible candidates for alignment. MWA considers three levels of similarity. The first one is exact word or lemma match, which is represented by the score of 1. The second level represents words that are not identical but are supposed to be semantically similar based on their appearance in lexical resources. Specifically, Paraphrase Database (Ganitkevitch et al., 2013), a large resource of lexical and phrasal paraphrases constructed using bilingual pivoting (Bannard and Callison-Burch, 2005) is employed.[4] Finally, any pair of different words that were not identified at the previous two levels are assigned zero score.

Contextual evidence is defined as the sum of word similarities of the words in context. A minimal-evidence assumption is applied which holds a single piece of contextual evidence as sufficient support for a potential alignment. Matching content word pairs are aligned even in the absence of context similarities. This is motivated by an empirical observation that more often than not content words are inherently sufficiently meaningful to be aligned even in the absence of contextual evidence when there are no competing word pairs.

Two types of context are taken into consideration: textual neighborhood (content words within (3,3) window) and syntactic neighborhood. The context is constituted by the head and dependent nodes in the dependency graph of the sentence (see the next Section for a detailed discussion of the representation of syntactic context). For textual neighborhood, context words are considered evidence for alignment if they were found to be lexically similar. In the case of syntactic context, context words are considered as evidence for alignment if they were found to be lexically similar and have the same or equivalent syntactic relations with the words to be aligned. Syntactic functions are defined as equivalent if they instantiate the same semantic relation. Sultan et al. (2014) have developed a mapping which defines the equivalence of dependency functions for four mayor lexical categories: verbs, nouns, adjectives and adverbs. For example, the dependency relation between subject and predicate in an active clause and by-agent and predicate in a passive clause are considered equivalent. We further use this mapping in the scoring stage in order to avoid penalizing acceptable variation at syntactic level (see Section 4.3.2).

As an illustration of the advantages of contextually informed alignment, consider

[4]MWA does not work with phrasal alignments. They use the largest (XXXL) of the PPDB's lexical paraphrase packages. The pairs of identical words or lemmas are removed and lemmatized forms of the remaining pairs are added. The number of word pairs remaining after this procedure is approx. 210K. Note in comparison, that the paraphrase database used by Meteor contains approx. 5M lexical and phrasal paraphrases (Denkowski and Lavie, 2011).

the alignment for the example from Table 4.1 above, generated by Meteor and MWA (Figure 4.1).[5] Alignments marked in green indicate aligned words that resulted from translating the same parts of the original sentence. These words have a similar function in the candidate and reference translations and are indeed indicative of sentence similarity. Alignments marked in yellow represent the words that correspond to the same parts of the source sentence, but play different roles in the translated sentences due to MT errors. For example, the direct object expressed in the source sentence by a nominalized adjective ("best") was incorrectly translated as an adjectival modifier of the subject ("judges") in the MT output. The match between those words, therefore, does not contribute to semantic similarity between candidate and reference sentences. Finally, alignments marked in red represent spurious matches between candidate and reference words that were generated from different source words. The matches of the MT words "the", "to" and "of" would increase the evaluation score for n-gram based metrics, while being completely meaningless with regards to candidate-reference similarity or the quality of MT output. MWA does not align function words unless they contain minimum context evidence (at least one pair of aligned content words in their context).[6]

Source:	Eso es lo que hemos hecho: hemos evitado una crisis presupuestaria recurriendo al artículo 272; se han mantenido las Perspectivas Financieras, pese a haber utilizado el instrumento de la flexibilidad.
Ref*:	*This is what we have done : we have avoided a budgetary crisis resorting to Article 272; the financial perspectives have been maintained, despite having used the flexibility instrument* .
MT:	That is what we have$_1$ done : we have$_2$ avoided a budgetary crisis recourse to Article 272 does the$_1$ financial perspectives have$_3$ been , despite having$_4$ used the$_2$ flexibility instrument .
Ref:	That is what we have$_1$ done : we have$_2$ avoided a budgetary crisis by going to Article 272 ; the$_1$ financial perspective has$_3$ been maintained , even though we have$_4$ used the$_2$ flexibility instrument .

Table 4.3: Example of ambiguity in the alignment between candidate and reference translations (WMT2007-Europarl dataset, Spanish–English translation, sentence 1603)

Furthermore, combining lexical similarity information with context evidence allows

[5] Alignments between punctuation marks, as well as non-aligned words are omitted for simplicity.

[6] In the case of Meteor this may have an opposite effect on the score. An increase in the number of lexical matches can be canceled by the increase in fragmentation penalty if the matching words are found in different positions in the sentences and the surrounding words are aligned. Spurious matches are thus treated as differences (i.e. errors) in word order. We consider that not aligning those words is more appropriate, given that they are often related to different parts of the source sentence.

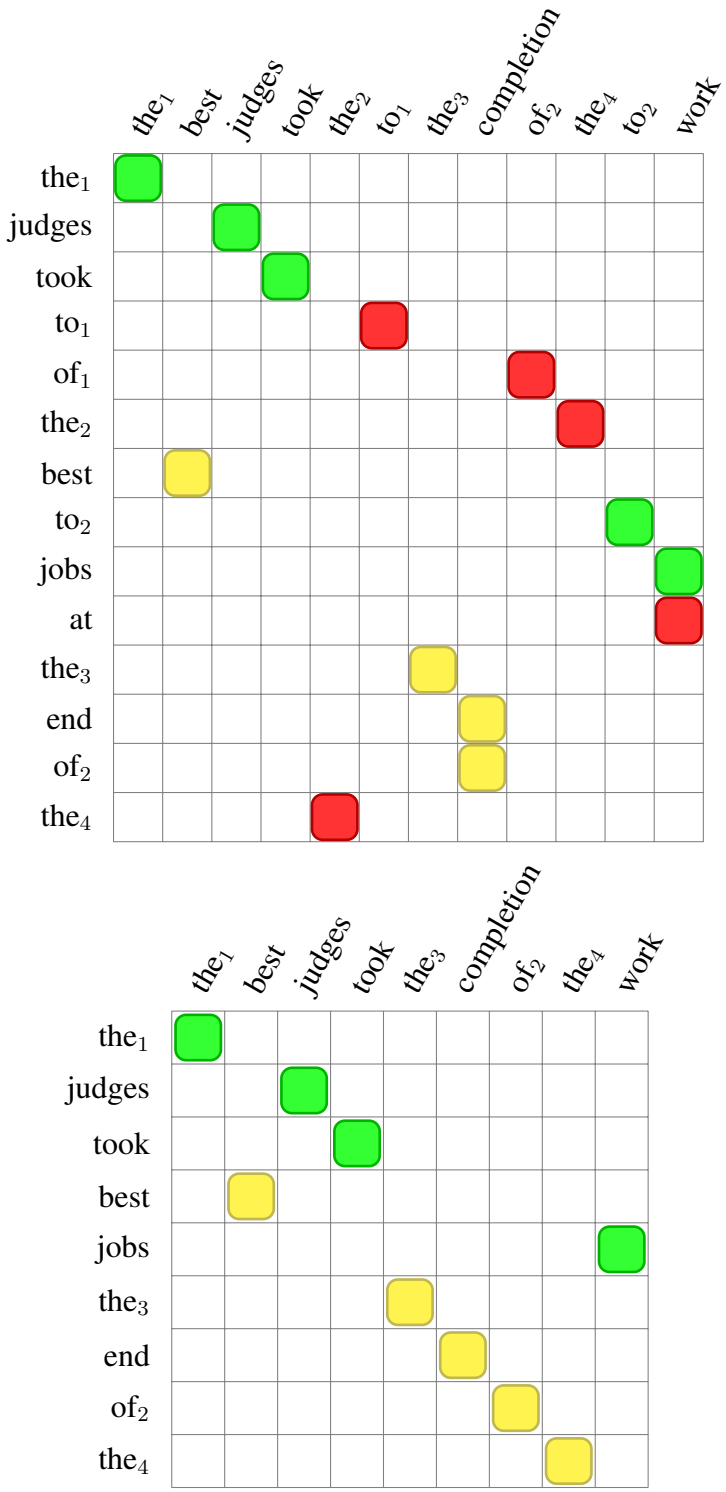


Figure 4.1: Meteor (top) and MWA (bottom) alignments for the example from Table 4.1

to discriminate between alignment candidates. Consider the example in Table 4.3 and the corresponding alignment matrices from Meteor and MWA (Figure 4.2). This example illustrates the third issue from (MacCartney et al., 2006). Meteor aligner prefers the alignment between the exactly matching verb forms $have_{3MT}$ and $have_{4REF}$, even though they occur in different context in the candidate and reference sentences and, in fact, correspond to different source words. By contrast, the strength of lexical matches is successfully combined with context evidence in MWA alignment and the correspondence between the words $have_{3MT}$ and has_{3REF} in the candidate and reference translations is correctly established. This is beneficial for MT evaluation since, in order to properly compare candidate and reference translations we are interested in establishing correspondence between translated words that come from the same original words, although they are not necessarily the most lexically similar or differ in their morphological forms. In this way, the type of translation error can be appropriately determined. The difference is not in the position of the verb, but in the verb forms used.[7]

Improving Alignment Lexical Coverage with Distributed Word Representations

Recall from Section 3.1.3 that lexico-semantic shifts in translation include modification (when source and target translation units are not straightforward equivalents of each other because of a difference in one or various aspects of their meaning), explicitation (when the target translation unit is lexically more specific than its source counterpart) and generalization (or implicitation, when the target translation unit is less specific than its source counterpart) (Cyrus, 2006; Ahrenberg and Merkel, 2000).

Therefore, lexical variation between different translation alternatives in the target language goes far beyond synonymy. Different translations of the same source sentence may contain hypernyms and hyponyms. Recall the following example from Chapter 1 repeated below (Table 4.4).

A more specific term “officers” is used in the reference translation, corresponding to the original word “people”, which is preserved in the MT output, resulting in a different, but perfectly acceptable translation.

In general terms, different translations of the same original sentence can contain contextual synonyms, i.e. words that are synonymous given certain context of use. As an illustration, consider the example in Table 4.5. The correspondence between the words “agreement” and “consent” can be easily established with the help of common lexical similarity resources such as WordNet. This is not the case, however, with the

[7] Although Meteor aligner (see Section 2.1.1) prefers alignments with a minimum number of crossing links, a) different types of matchers are applied one after the other imposing priority b) Meteor does not use syntactic context, therefore acceptable and non-acceptable differences in the word order are treated in the same way. The latest version of Meteor aligner (Denkowski and Lavie, 2014) considers all possible matches in a single alignment stage. However, if a pair of words constitutes an exact match, they are not considered for alignment using stemming or synonym matchers.

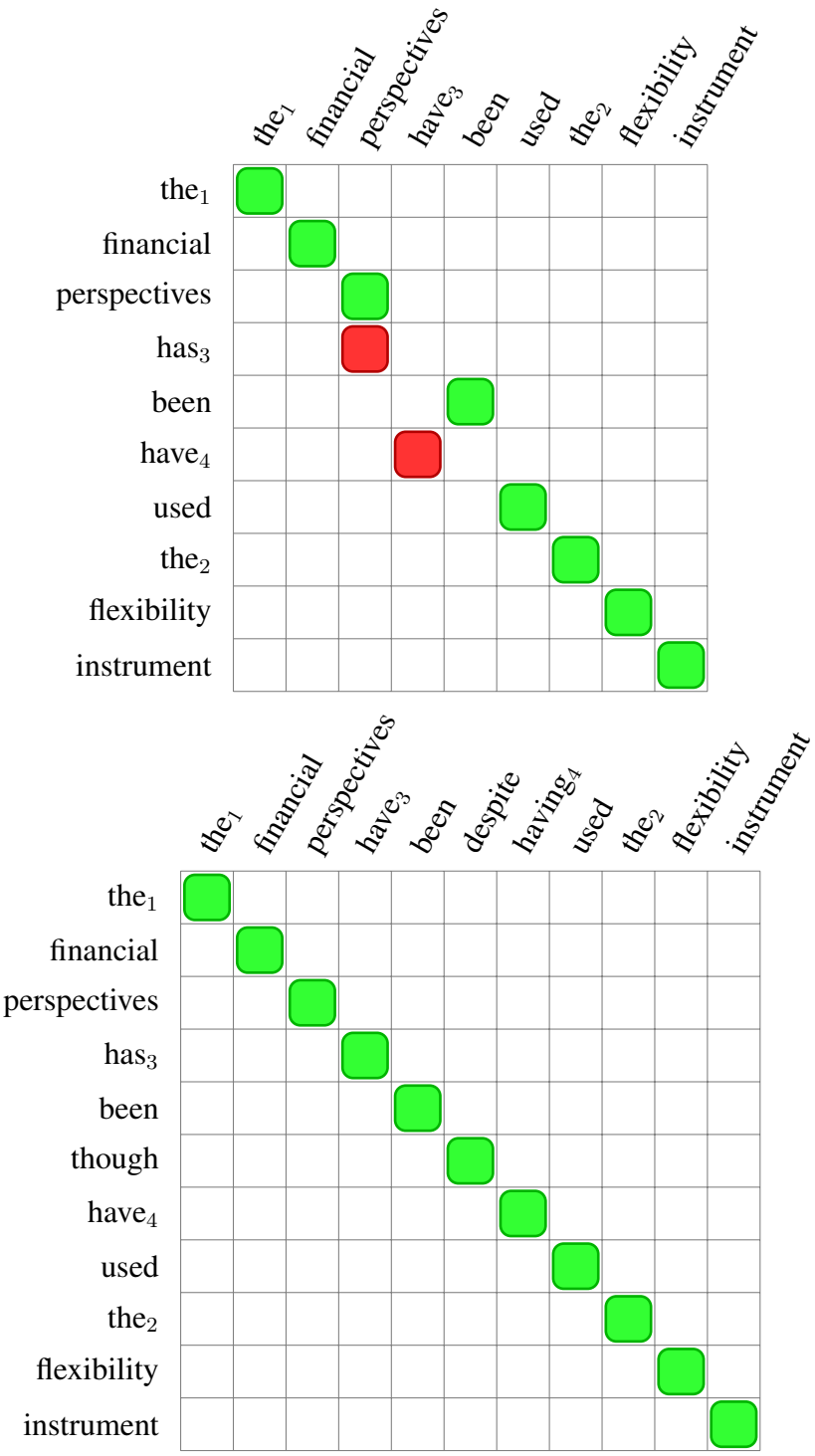


Figure 4.2: Meteor (top) and MWA (bottom) alignments for the example from Table 4.3

Source:	Ламб сказал диспетчеру, что полиции необходимо послать людей к нему домой.
Ref*:	Lamb told dispatcher that police needs to send people to him home.
Ref:	Lamb told the dispatcher that police needed to send **officers** over to his home.
MT:	Lamb told the dispatcher that the police needs to send **people** to him home.

Table 4.4: Example of hypernymy in reference-based MT evaluation (WMT16 dataset, Russian–English translation, sentence 688)

Source:	Tengo entendido que el Consejo también ha **dado** en principio su consentimiento.
Ref:	I understand that the Council has also **signalled** its agreement in principle.
MT:	I understand that the Council has also **given** its consent in principle.

Table 4.5: Example of contextual synonymy in reference-based MT evaluation (WMT2007-Europarl dataset, Spanish–English translation, sentence 1520)

words “signalled” and “given”, which can be considered semantically equivalent only given the equivalence of their contexts.

In order to increase the number of cases like the above to be recognized as acceptable variation in MT evaluation, we propose using distributed word representations with additional constraints based on the difference in their contexts. Distributed word representations are grounded on the distributional hypothesis (Harris, 1954) that states that semantic similarity between two words can be modeled as a function of the degree of overlap between their contexts. In this framework, words are represented as vectors in which each component is a measure of association between the word and a particular context. The similarity between two given words is then computed using some distance measure on the corresponding vectors (e.g. cosine similarity).

Distributional semantic models (Baroni and Lenci, 2010) have been shown to perform well across a variety of lexical similarity tasks, including hypernymy detection (Shwartz et al., 2017). In particular, representing words as dense vectors derived by training methods inspired from neural-network language modeling (commonly referred to as word embeddings) proposed by Mikolov et al. (2013) have been shown to outperform previous approaches (Baroni et al., 2014). We integrated a lexical similarity component based on word embeddings in the MWA alignment system in order to increase lexical coverage.

Whereas WordNet and paraphrase databases are commonly used in MT evaluation for dealing with lexical variation (Snover et al. (2009b); Denkowski and Lavie (2010a), *inter alia*), to the best of our knowledge, distributional similarity has not yet been exploited. Following our work in Fomicheva et al. (2015b), Servan et al. (2016) explored the use of different lexical resources for the purposes of MT evaluation including word embeddings by means of integrating this component to Meteor. In Section 4.4.1 we compare our approach with the results reported in Servan et al. (2016) and show that our approach outperforms their proposal by a large margin, due to the use of additional context constraints that reduce the noise introduced by this component while still taking advantage of the increased lexical coverage.

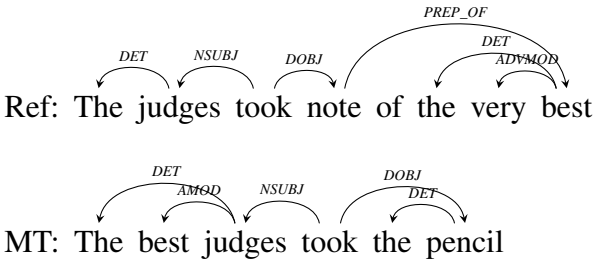
Previous work on neural word embeddings takes the contexts of a word to be its linear context, i.e. preceding and following words, typically in a window of k tokens to each side. In the work of Levy and Goldberg (2014), the skip-gram model from Mikolov et al. (2013) is generalized to move from linear bag-of-words contexts to syntactic contexts that are derived from automatically produced dependency parse-trees. The authors showed that the different kinds of contexts produce noticeably different embeddings, and induce different word similarities. In particular, the bag-of-words nature of the contexts in the "original" skip-gram model yields broad topical similarities, while the dependency-based contexts yield more functional similarities of a cohyponym nature (*cf.* the distinction between domain similarity versus functional similarity in Turney (2012)). In other words, when measuring similarity to a target word, bag-of-words models find words that associate with the target word, while dependency-based models find words that behave like the target word. As we are interested in paradigmatic relationships between words, we use dependency-based word embeddings in our work. Section 4.3.1 describes in detail how we use vector representations for candidate-reference alignment.

4.2 Syntactic Word Context

In the previous Section we discussed monolingual word alignment as a means for establishing correspondences between candidate and reference words. After an alignment between candidate and reference translations is defined, MT evaluation can be conducted based on the number of matching words and the differences in their syntactic contexts. In this Section we turn to the use of syntactic context of the aligned words in order to test the validity of the established word correspondences. Here we provide a general motivation and discussion, while Section 4.3.2 presents how the difference in the syntactic contexts of the words are used for scoring MT output.

As we have seen in Section 4.1, a match between the words that occur in different contexts barely contributes to sentence similarity. Therefore, we can assume that matching candidate-reference words occurring in different syntactic contexts are indicative of

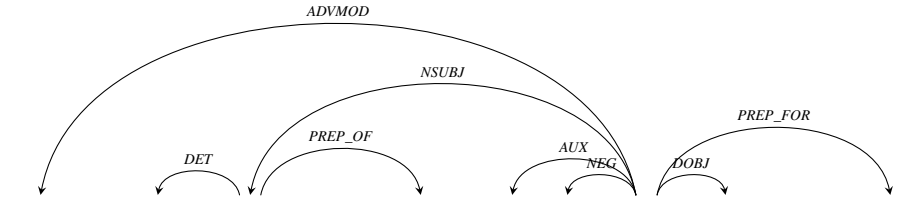
word order errors, wrong lexical choice, wrong choice of function words, morphological errors, etc. Consider the dependency syntactic structure of a part of the example from the previous Section (Table 4.1):



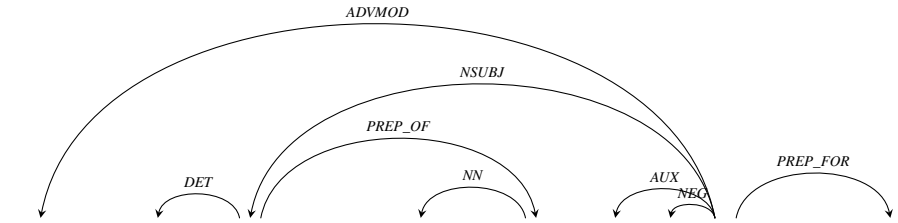
The match of the word "best" does not contribute to semantic similarity between the two sentences. In the reference translation "best" is the prepositional object of "took", whereas in the MT output it is an adjectival modifier of the subject, due to the mishandling of the original word order. Clearly, the lexical match in this case should not have the same positive effect on the evaluation score as in case it had the same function in the two sentences.

Two types of syntactic representations are traditionally used in NLP, constituent-based trees and dependency trees. The latter are preferred for sentence similarity tasks, as they conveniently leverage lexical and syntactic information. Following the modeling of syntactic context evidence in MWA (see Section 4.1), we use dependency representation to extract the syntactic context of the words. Specifically, for each aligned word its syntactic context is defined as its head and dependent nodes in the dependency graph. Stanford typed dependencies from De Marneffe and Manning (2008) are used as syntactic representation which is obtained from the Stanford dependency parser (De Marneffe et al., 2006). A collapsed representation of dependencies is chosen, where dependencies involving prepositions, conjuncts, as well as information about the referent of relative clauses are collapsed to get direct dependencies between content words. In our view, this is beneficial for MT evaluation task, as it allows for better generalization.

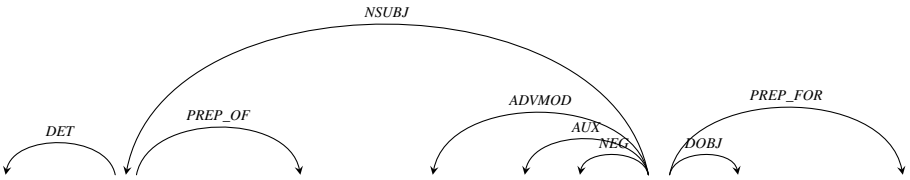
N-gram based metrics also take context into account. The scores of n-gram metrics such as BLEU (Papineni et al., 2002) depend on the number of matching sequences of words if $n > 1$. Fragmentation penalty from Meteor (Denkowski and Lavie, 2014) aimed at explicitly penalizing the differences in word order, downscales the evaluation score when matching words occur in a different context (i.e. position) in the candidate and reference translations. The advantage of using a syntactic representation over surface linear context is twofold. On the one hand, it allows to abstract away from the word sequences and avoid penalizing acceptable differences in word order. On the other hand, it allows to automatically establish the importance of certain differences by checking to what extent they affect the syntactic parse of the sentence. Consider the following hypothetical case as an illustration.



(1) However, the Chamber of Deputies did not release her for prosecution.



(2) However, the Chamber of Deputies release did not her for prosecution.



(3) The Chamber of Deputies, however, did not release her for prosecution.

Word	Context	Number
However	(release, ADVMOD)	1
the	(Chamber, DET)	1
Chamber	(the, DET), (Deputies, PREP_OF), (release, NSUBJ)	3
Deputies	(Chamber, PREP_OF)	1
did	(release, AUX)	1
not	(release, NEG)	1
release	(ROOT, ROOT), (However, ADVMOD), (Chamber, NSUBJ), (did, AUX), (not, NEG), (her, DOBJ), (prosecution, PREP_FOR)	7
her	(release, DOBJ)	1
prosecution	(release, PREP_FOR)	1

Table 4.6: Definition of syntactic context based on dependency representation

Sentence (2) is agrammatical as it contains an unacceptable alternation of word order. In sentence (3), on the other hand, the change in word order does not lead to agrammaticality, as English admits various positions for adverbial modifiers. Taking linear word sequences as the basis for comparison, sentences (2) and (3) are equally different from sentence (1). The change in the position of the adverbial modifier does

not affect the dependency parse, whereas the parse generated for sentence (3) is substantially different. In this way, comparing syntactic representations allows to distinguish between acceptable and non-acceptable differences.

Table 4.6 shows an example following the definition of syntactic context that we use in this work. The last column in Table 4.6 shows the number of context words corresponding to each word in the sentence.[8] We note that this information can be used as an additional feature to characterize the impact of the candidate-reference differences on the perceived quality of the MT output. We do not dispose of any hard evidence that the number of dependents of mistranslated words is related to the impact of the errors on the perception of MT quality.[9] However, we hypothesize that such relation may exist and prove this indirectly by attaining an improved correlation with human judgments.[10]

We use a syntactic representation of the context instead of the word window, thus placing ourselves in the line of work that integrates linguistic information in the MT evaluation, which has been shown to ameliorate the limitations of n-gram matching (Liu and Gildea, 2005). This allows to better capture the well-formedness of the MT output. N-gram-based evaluation metrics are unable to properly penalize statistical MT errors. When various long matches between candidate and reference phrases are found, the score gets relatively high, while the sentence is ungrammatical as separate unrelated phrases are stacked together. Syntactic representation alleviates this problem. On the other hand, as a more abstract representation, it helps to avoid penalizing acceptable differences in word order. Generally speaking, using alternative sources of information avoids the circularity of evaluation, i.e. making the same assumptions and using the same resources as the ones used by the MT systems under evaluation, the metrics fail to detect the failures of those systems.

A well-known concern regarding the use of syntactic information for MT evaluation is that syntactic parsers are not designed to process agrammatical sentences. Modern statistical parsers are likely to assign well-formed structure even to grammatically unacceptable sentences. However, as pointed out by Liu and Gildea (2005)

> in MT evaluation we are looking for similarities between pairs of parse trees rather than at features of a single tree. This means that the syntax-based evaluation measures can succeed even when the tree structure for a poor hypothesis looks reasonable on its own, as long as it is sufficiently

[8] The head node of the predicate "release" is the ROOT, not shown in the representation for simplicity.

[9] Some work has been developed investigating the influence of different types of translation errors on the perception of MT quality (Federico et al., 2014). In our view, this is an interesting research topic that needs to be further investigated.

[10] As an artifact of how BLEU score is calculated, it is, in fact, affected by different matches in a different way. But the effect depends on the position of mismatched words. The decrease in the score is more pronounced if mismatches occur in the middle of the sentence, since they affect a higher number of n-grams. This effect hardly has any motivation regarding human perception of MT quality.

distinct from the structures used in the references. (Liu and Gildea (2005, p.32))

In Chapter 3 we established that linguistic variation in translation goes far beyond the differences in lexical choice. Specifically, we discussed syntactic shifts in human translation. We established that for the purposes of MT evaluation, it is reasonable to treat syntactic constructions that convey similar semantic relations as equivalent. We used the equivalence between a set of syntactic constructions to generate additional reference translations. In this Chapter we address acceptable syntactic variation directly when computing similarity between candidate and reference translations. In order to achieve this, we take advantage of the system of equivalent dependency types developed by Sultan et al. (2014) for the purposes of computing context evidence for monolingual word alignment (see Section 4.1). Syntactic functions are defined as equivalent if they instantiate the same semantic relation. Sultan et al. (2014) have developed a mapping which defines the equivalence of dependency functions for four major lexical categories: verbs, nouns, adjectives and adverbs. Some examples of such functions are: possession modifier and noun compound modifier, indirect object and prepositional modifier, relative clause modifier and reduced non-finite verbal modifier, nominal subject of an active clause and by-agent in a passive clause. See Sultan et al. (2014) for a complete list of equivalent syntactic relations. The example in Table 4.7 illustrates the idea of the equivalence of syntactic contexts. In this example MT1 is a perfectly acceptable translation, whereas MT2 contains an error concerning the relation between the words "voter" and "Obama". Using the mapping between equivalent dependency functions, it can be established that the syntactic contexts of the aligned words in MT1 are equal or equivalent. Thus, the *prep_for* relation in the candidate translation is equivalent to the noun compound modifier relation *nn* in the reference and the *prep_of* label in the candidate corresponds to the possession modifier *poss* in the reference. By contrast, MT2 contains differences in the syntactic contexts of the words "voter" and "Obama", which constitute a translation error.

4.3 UPF-Cobalt Metric

We implemented the approach described in the previous Sections as a full-fledged evaluation metric, which we called UPF-Cobalt (Fomicheva and Bel, 2016). Our metric proceeds in two stages. In the first stage, an alignment between candidate and reference translation is established. In this stage the metric uses MWA aligner because of the benefits discussed in Section 4.1. In the second stage, the score for the MT sentences is computed based on the characteristics of candidate-reference matches. Crucially, in this stage, besides considering the type of the match (e.g. exact, stems, lemmas or synonym matches), the difference between the syntactic contexts of the matching words is com-

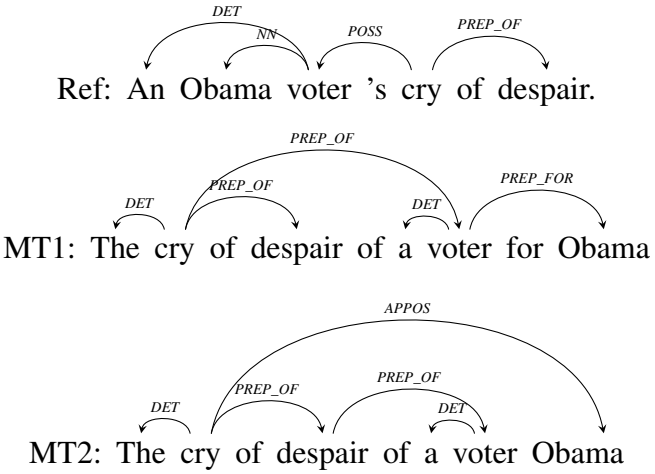


Table 4.7: Example of equivalent syntactic contexts in candidate and reference translations (WMT2014 dataset, French–English translation, sentence 1061)

puted and integrated in the overall sentence-level score. The metric and the code are freely available online at `http://github.com/mfomicheva/upf-cobalt`.

4.3.1 Alignment

As mentioned above, we use MWA in this stage, but in principle any other aligner can be used (see Section 4.5.2 for the results of the experiments with Meteor Aligner). Recall from Section 4.1 that MWA relies on stem, lemma matches and paraphrase matches (Ganitkevitch et al., 2013) to allow for the alignment of similar words. To increase lexical coverage, we integrate two additional levels to the MWA's lexical similarity component. In addition to the Paraphrase Database, we employ WordNet synonyms (Fellbaum, 1998) and distributed word representations (Mikolov et al., 2013). As discussed in Section 4.1, the combination of lexical-semantic resources and word vector representations has barely been explored in previous work. To the best of our knowledge, ours is the first work to apply a measure of distributional similarity to the task of reference-based MT evaluation.

WordNet is used as in other works on sentence similarity, paraphrase detection or MT evaluation, i.e. words are considered similar if they are members of the same WordNet synset. Regarding the distributional similarity component, words are required to meet several constraints in order to be considered as potential candidates for alignment. First, they must have a cosine similarity between the corresponding vector representations higher than a threshold.[11] The threshold was established heuristically. We found

[11]For the vector representations we use an existing word embeddings resource described in Levy and

that the threshold of 0.25 provides a good trade-off between precision and recall if an additional context constraint is introduced. Second, the words are required to have at least one pair of exactly matching content words in their contexts. Recall from Section 4.1 that MWA aligns content words even if no context evidence was found, as content word matches are considered sufficiently meaningful. We modify this principle for the distributional similarity component. It must be noted that MT evaluation is different from the tasks of sentence similarity or paraphrase detection in that MT output may contain translation errors to the extent that the matches between the words may be completely meaningless towards sentence similarity. This fact, combined with the fact that cosine similarity over vector representations is a noisy source of information motivates the introduction of an additional contextual restriction aimed at reducing the number of noisy correspondences established in the alignment stage. Along the same lines, an additional restriction concerns the alignment of function words. Function words can be aligned based on the distributional similarity component only if they belong to the same grammatical category, or the grammatical categories involved are a pronoun and a noun. This latter exception is introduced to account for the case of pronominalization and de-pronominalization shifts discussed in detail in Chapter 3. If the MT contains a noun and the reference contains a pronoun occurring in the same context, it is reasonable to assume that the difference is due to a translation shift in the reference.

Recall from Section 4.1 that for each pair of words that can potentially be aligned, the MWA alignment score is computed based on their lexical similarity and the context evidence. The set of lexical weights for computing alignment scores in MWA are as follows: exact, lemma or stem match - 1.0, paraphrase match - 0.9 and zero otherwise. After adding WordNet synonyms and distributional similarity as additional levels of lexical similarity, we set lexical weights for computing alignment scores heuristically, depending on the strength of the match and the level of noise in the resource:

- same word forms - 1.0
- same stem or lemma- 0.9
- same WordNet synset - 0.8
- paraphrase match - 0.6
- distributional similarity - 0.5

4.3.2 Scoring

Once the alignment with the reference sentences has been established, MT outputs can be scored based on the similarities and differences detected. We follow the method proposed by Denkowski and Lavie (2014) in computing the sentence-level score as a weighted combination of precision and recall over the number of aligned word pairs.

Goldberg (2014) and freely available for download from `https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/`.

However, the score for each pair of aligned words is computed as a combination of the lexical match score and a context penalty based on the number of different words in their syntactic contexts. As will be shown in what follows, context penalty quantifies the importance of local candidate-reference differences for the sentence-level score.

Sentence-Level Score

Recall from Section 2.1.1 that Meteor's sentence-level score is based on a weighted combination (F_{mean}) of Precision (P) and Recall (R) over different types of lexical matches (m) (exact, stem, synonym and paraphrases) between candidate (t) and reference (r) words. Precision and Recall are weighted using the match weights and the content-function words weight (δ).[12] For clarity, we reproduce here the formulas from Section 2.1.1.

$$F_{mean} = \frac{P \times R}{\alpha \times P + (1 - \alpha) \times R} \tag{4.1}$$

$$P = \frac{\sum (\delta \times m_i(t_c) + (1 - \delta) \times m_i(t_f))}{\delta \times |t_c| + (1 - \delta) \times |h_f|} \tag{4.2}$$

$$R = \frac{\sum (\delta \times m_i(r_c) + (1 - \delta) \times m_i(r_f))}{\delta \times |r_c| + (1 - \delta) \times |r_f|} \tag{4.3}$$

UPF-Cobalt score is computed using the equations 4.1-4.3. Parameters $\alpha = 0.85$ and $\delta = 0.75$ from the default version of Meteor are used. Instead of the lexical match weights of the aligned candidate and reference words we plug in the numerator of the equations 4.2 and 4.3 the word-level scores ranging from 0 to 1 and obtained as a combination of lexical similarity ($LexSim$) and context penalty (Pen):

$$score(t, r) = LexSim(t, r) - Pen(t, r) \tag{4.4}$$

Lexical Similarity

Recall from the previous Section that words can be aligned if they have the same surface form, if they have the same stem or lemma, if they were found in the same WordNet synset, if they were found in the paraphrase database or if the cosine similarity between their distributed representations is higher than the threshold of 0.25. We use the same weights for alignment and for scoring.[13] It should be noted that the function of lexical

[12] A stopwords list from the NLTK toolkit (Bird, 2006) and a punctuation list from MWA are used to detect function words. Using Meteor's frequency based function word list does not produce any change in the results.

[13] The words aligned using the named entities component of MWA are also assigned a value of 0.5.

weights is different when establishing candidate-reference alignment and when computing the score for candidate translation. In the first case, the weights are needed to choose between candidate pairs for word alignment. Thus, assigning a higher weight to exact word matches, they are given preference compared to the matches identified using other lexical resources. In the case of scoring, lexical weights serve to describe the difference between the aligned words and how this difference will affect the sentence-level score. However, as the alignment weights from previous Section are based on the closeness of the match (exact vs. stem) and on the level of noise in the lexical resource used for alignment (synonym vs. distributional similarity), it is reasonable to use the same set of weights for alignment and for scoring. We experimented with optimizing the weights for different types of lexical similarity, as well as for the classes of dependency functions discussed below, using genetic algorithms (Mitchell, 1998). However, the optimization gave approximately the same values, showing that our intuitions were essentially correct.

Note also that by contrast to the match scores typically used in MT evaluation based on the accuracy of the lexical resources, cosine similarity over distributed word representations genuinely measures the similarity between the words. We experimented with (a) substituting all the lexical resources with distributional similarity and (b) using the cosine value directly instead of the heuristically established weight. However, this resulted in a significant decrease in the correlation, showing that cosine similarity over vector representations is not suitable for reference-based MT evaluation.[14] Aligning words using distributional similarity results in noisy alignments but the noise is minimized by the contextual restrictions and by the context penalty applied if the contexts of the words are different, as described below.

Context Penalty

A context penalty is applied at word level to identify cases where the words are aligned (i.e. lexically similar) but play different roles in the sentences and, therefore, should contribute less to the sentence-level evaluation score. Thus, for each pair of aligned words, the words that constitute their syntactic contexts are compared. Recall from Section 4.2 that the syntactic context of a word is defined as its head and dependent nodes in a dependency graph. Both the context words themselves and their dependency labels are compared.

[14]Recent work using neural networks for MT evaluation achieves reasonable performance only in combination with other evaluation metrics (Gupta et al., 2015; Guzmán et al., 2016). The main reasons seems to be the insufficient amount and quality of training data. However, the same strategy works fairly well for sentence similarity task (Tai et al., 2015). The main difference between sentence similarity and MT evaluation is the presence of MT errors that individually and in combination produce a different effect on the sentence-level score. One of the reasons why neural-based approaches to MT evaluation have not been able to succeed is the lack of appropriate treatment of MT errors. We leave this as an interesting direction for future work.

The following issues are taken into consideration when measuring contextual differences. First, to account for the possible equivalence of certain syntactic relations we use the equivalent dependency types mapping proposed by Sultan et al. (2014). As we have discussed in Chapter 3, syntactic variation is a regular source of differences between human translation and MT. By taking it into consideration, we avoid penalizing perfectly acceptable MT outputs that contain different syntactic structures but that are semantically similar to the reference. Second, mistranslating the words with argument functions (subject, direct object, prepositional object, etc.) changes the context to a greater extent than dropping a determiner or an adjunct. Therefore, context words are assigned different weights depending on the relative importance of their syntactic functions. Finally, the number of context words is taken into account assuming that a translation error involving a word with more syntactic dependents has a higher impact on the MT quality.

For each pair of aligned words, t in the candidate translation and r in the reference translation, we define the context penalty as follows:

$$\begin{aligned} CP(t,r) &= \frac{\sum_{1..i} w(C_i^*)}{\sum_{1..i} w(C_i)} \ln\left(\sum_{1..i} w(C_i) + 1\right) \\ Pen(t,r) &= \frac{2}{1 + e^{-CP(t,r)}} - 1 \end{aligned} \tag{4.5}$$

Where CP stands for context penalty, C refers to the words that belong to the syntactic context of the word under consideration and C_i^* refers to the context words that are **not** the same or equivalent. For the words to be equivalent two conditions are required to be met: a) they must be aligned and b) they must be found in the same or equivalent syntactic relation with the word under consideration. For each pair of aligned words, CP is calculated using the candidate context and the reference context. A weighted average of these values is used as the final CP value.

The weights w that reflect the relative importance of the dependency functions of the context words are defined as follows:

- argument/complement functions - 1.0
- modifier functions - 0.8
- specifier functions - 0.2

The number of context words is taken into consideration assuming that the higher the number of syntactic dependents a word has, the higher should be the impact of a candidate-reference difference involving this word on the sentence-level score. We use the natural logarithm of the weighted count of context words, since this impact saturates above a threshold. Thus, a context difference receives a higher value when the number of context words is high (it is not the same translating zero words out of one and zero words out of ten), while limiting the increase if the number of context words continues

to grow (the difference between translating six words out of eight and eight words out of ten is less relevant). To obtain the final value for context penalty (Pen), CP is normalized from 0 to 0.5 using the logarithmic function.

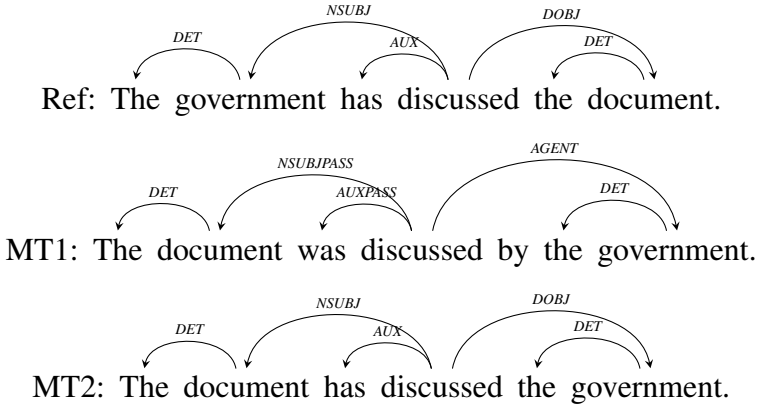


Table 4.8: Example for the comparison of UPF-Cobalt with state-of-the-art evaluation metrics (WMT2014 dataset, Czech–English translation, sentence 2272)

To appreciate the advantages of the method proposed, the example in Table 4.8 provides a qualitative comparison of the performance of UPF-Cobalt and other state-of-the-art evaluation metrics. The scores assigned to MT1 and MT2 translation candidates are shown in Table 4.9. Meteor and BLEU are the usual benchmark systems. The well-known head-word chain method from Liu and Gildea (2005) discussed in Section 2.1.3 (HWCM) (Liu and Gildea, 2005) is also presented. As suggested by Giménez and Màrquez (2010), HWCM can be used with different representations, w stands for words, c stands for categories and r stands for dependency relations. Meteor and HWCM serve as nice points of comparison, as UPF-Cobalt is similar to Meteor in that it is an alignment-based system using similar lexical resources and to HWCM in the sense that it uses the information extracted from the dependency parse. Also, the scores from some of the metrics that also leverage different types of linguistic information (VERTA and DiscoTK) from the WMT14 dataset, from where this example was taken, are compared.

MT1 is assigned a lower score than MT2 by all of the metrics in Table 4.9 except for UPF-Cobalt, due to the change in surface word order. UPF-Cobalt correctly assigns a high score to this sentence. All the content words are aligned and no context penalty is applied, since the syntactic contexts of the aligned words are equivalent. Thus, *agent* relation in the candidate translation is equivalent to nominal subject relation (*nsubj*) in the reference, and subject of a passive clause (*nsubjpass*) in the candidate corresponds to the direct object (*dobj*) in the reference.

By contrast, the other metrics assign a higher score to MT2 because of a matching auxiliary verb which in this case is not indicative of candidate-reference seman-

tic similarity. MT2 receives a much lower score from UPF-Cobalt. Although all the content words are matched they occur in different contexts and receive a high context penalty (0.90 for the main verb "discussed" and 0.80 for the arguments "government" and "documents"). Thus, UPF-Cobalt is capable of distinguishing the use of equivalent constructions (active/passive alternation) from translation errors.

Note that, in UPF-Cobalt, when comparing the syntactic contexts of the aligned words, both the words and the dependency relations are taken into consideration. This is different from the HWCM metric. Head-word chain method compares word n-grams extracted from the dependency path instead of word n-grams extracted from a linear word sequence. The score is then computed based on the number of n-gram matches in a BLEU-style manner. This does not consider neither fuzzy matches between words, nor their syntactic functions. Note that MT2 receives a maximum score from all the three types of representation. This illustrates very well the point that lexical and syntactic information needs to be combined at word level, not at sentence level.

Metric	MT1	MT2
UPF-Cobalt	0.916	0.6525
HWCM_w-4	0.5385	1.0
HWCM_c-4	0.5714	1.0
HWCM_r-4	0.375	1.0
VERTA-EQ	0.6726	0.9167
VERTA-W	0.6995	0.9367
DiscoTK$_{PARTY}$	0.495	0.7209
Meteor	0.4077	0.4635
BLEU	0.1729	0.3593

Table 4.9: Metric scores for the example in Table 4.8

We note that the word-level context penalty captures the propagation of translation errors. If the mistranslated word has many syntactic dependents all of them receive a context penalty, which strongly affects the score at sentence level. By contrast, if the error involves a word that has few syntactic dependents its impact will be low. Thus, the redundancy present in the scoring method is an advantage. As context penalty is calculated for each aligned word in the sentence, the score decreases depending on the number of words affected by the error.

The above formulation also allows us to meet the requirement discussed in the beginning of this Chapter. Lexical matches increase the sentence-level evaluation score depending on the number of different words in their contexts. Conversely, candidate-reference differences decrease the evaluation score depending on how many matching words they affect. Thus, the impact of different words is estimated in terms of the

number of matching words in their context.

4.4 Meta-Evaluation

The ultimate goal of automatic evaluation metrics is to emulate manual assessment of MT quality. Therefore, the performance of the metrics is typically evaluated by computing the correlation between the metrics scores and human judgments. We conduct experiments with different types of manual annotation to show the robustness of the method in varying evaluation settings (for a detailed description of different types of manual evaluation, see Section 2.2). We compare our approach to a set of state-of-the-art evaluation metrics, showing that comparing local syntactic contexts of the matching candidate-reference words results in a more accurate MT evaluation.

4.4.1 Ranking Judgments: WMT14-WMT16 Datasets

We start with ranking human judgments (also known as preference judgments). This evaluation setting involves comparing the quality of the outputs of different MT systems for the same source sentence. Specifically, we use the ranking data from Annual Workshops on Statistical Machine Translation (WMT). The datasets consist of source texts in the news domain, human reference translations and the outputs from the participating MT systems, for various language pairs (see Section 2.2.2).

Following the practice introduced at WMT, we assess the performance of the metrics in terms of the Kendall rank correlation coefficient based on the number of concordant and discordant pairwise comparisons. A concordant pair is a pair of two translations of the same segment in which the comparison of human ranks agree with the comparison of the metric's scores. A discordant pair is a pair in which the comparison of human ranks disagrees with the metric's comparison. Specifically, the definition of Kendall τ presented in Macháček and Bojar (2014b) is used. This was the official measure for the WMT Metrics Task starting from 2014. In this definition, human ties are ignored.

The focus of our work is on sentence-level evaluation. Simple n-gram-based methods can attain a relatively high correlation at system level. The number of matching words and phrases in system and human translations is a reasonably good predictor of quality if averaged over many sentences. Note that even just counting the number of MT outputs that exactly match the reference over thousands of sentences is a reasonably good measure of quality. Thus, system-level automatic evaluation can be largely considered a solved problem. At the WMT16 metrics task the average Pearson correlation for into-English translation at system level was 0.94 with baseline BLEU system achieving 0.92). The real challenge for automatic evaluation, therefore, lies in the sentence-level task.

	Metric	WMT13	WMT14	WMT15	WMT16
	UPF-Cobalt	0.273	0.367	0.422	0.385
I	BLEU	0.197	0.285	0.349	0.294
	SimpBLEU-recall	0.215	-	-	-
	SimpBLEU-prec	0.211	-	-	-
	ChrF3	-	-	0.407	0.364
II	Meteor	0.264	0.354	0.407	0.367
	Meteor Vector	-	0.328	-	-
	Depref-Align	0.238	-	-	-
	Depref-Exact	0.234	-	-	-
	RedCombSent	-	0.356	-	-
	RedCombSysSent	-	0.356	-	-
	VERTA	-	0.337	0.387	-
III	BEER	-	0.362	0.421	0.368
	BEER_Treepel	-	-	0.429	-
	DiscoTK-Party-Tuned	-	0.386	-	-
	DPMFcomb	-	-	0.447	0.415
	RATATOUILLE	-	-	0.425	-

Table 4.10: Sentence-level evaluation results for WMT13-16 datasets in terms of Kendall rank correlation coefficient (τ)

Table 4.10[15] summarizes the results for the data from WMT metrics tasks of different years (2013-2016).[16] The results are averaged over the available into-English language pairs for each dataset (see Appendix A for detailed results). In addition to UPF-Cobalt, we present the results for the 5 best performing systems participating in the WMT Metrics Task for each year. The metrics are grouped according to the underlying evaluation strategy. In Appendix A we include empirical confidence intervals computed using the bootstrap resampling method as proposed in Macháček and Bojar (2014b). However, it has recently been suggested that, since Kendall's τ is used in a non-standard way (we do not have a single overall ranking of translations, but rather rankings of sets of 5 translations), the accuracy of confidence intervals computed in this way is difficult to verify (Bojar et al., 2016). Unfortunately, no alternative method for computing confidence intervals in this setting has yet been defined.

Group I includes string-based evaluation metrics that do not use any additional information other than candidate and reference sentences. ChrF3 (Popovic, 2015) calculates a simple F-score combination of the precision and recall of character n-grams of length

[15]The results reported for the WMT13 dataset differ from the results published in Macháček and Bojar (2014a), since they computed Kendall τ using a different method that was later proven to be inaccurate (see discussion in Macháček and Bojar (2014b)).

[16]We participated in the WMT15 and WMT16 Metrics Tasks. Results are reported in Fomicheva et al. (2015b, 2016).

6. The F-score is calculated with $\beta = 3$, giving triple weight to recall. Character n-grams allow to reduce penalizing morphological variation, which is why this metric outperforms BLEU by a large margin. SimpBLEU-recall and SimpBLEU-prec (Song et al., 2013) are unigram recall and precision without brevity penalty. These metrics address some of the well-known limitations of BLEU at sentence level. First, using unigram matches addresses the problem of data sparseness that affects longer n-grams. Second, a genuine recall measure is far more informative than brevity penalty (Banerjee and Lavie, 2005).

Group II contains the metrics that make use of lexical resources and syntactic information. Meteor (Denkowski and Lavie, 2011) employs WordNet synonyms and a paraphrase database in order to account for acceptable variation at lexical level, and a fragmentation penalty to explicitly penalize differences in word order.[17] In line with our work presented in Fomicheva et al. (2015b), Meteor Vector (Servan et al., 2016) employs word embeddings to improve candidate-reference alignment. Meteor's aligner is augmented with a matching module based on distributed word representations. Words with cosine similarity over word embeddings higher than a threshold are matched.[18] Interestingly, the reported oracle threshold obtained empirically on the WMT14 data is 0.86. Our proposal of using a lower threshold with context constraints outperforms this method by a large margin. Depref-Align and Depref-Exact (Wu et al., 2013) compute an F-score using string-based word n-grams of the candidate translation and n-grams extracted from the syntactic dependency of the reference translation. Dependency-based n-grams are extracted using the head-word chains method proposed by Liu and Gildea (2005). The main idea of the approach is to take the syntactic aspect into account avoiding the negative impact of parsing errors resulting from processing ill-formed MT output. RedCombSent and RedCombSysSent (Wu et al., 2014) improves on the metrics from the previous year by introducing additional lexical resources and tuning the weights for different types of lexical matches and n-gram lengths. VERTA (Comelles et al., 2012) computes candidate-reference similarity at lexical, morphological and syntactic levels by matching word, word with Part-Of-Speech (POS) tags and words with dependency labels, respectively. The final score is a weighted combination of the similarity scores obtained with different types of linguistic representation.

Group III includes feature-based approaches. All the metrics use learn-to-rank approach to tune the weight of the feature using ranking data. BEER (Stanojevic and Sima'an, 2014) uses character-based n-grams and permutation trees. BEER_Treepel (Stanojevic and Sima'an, 2015) includes features checking the match of each type of

[17] In the column "WMT13" in Table 4.10 the results obtained by Meteor 1.3 that participated in WMT13 Metrics Task (Denkowski and Lavie, 2011) are reported. In all the other experiments we report the results obtained with the latest version of Meteor (1.5) (Denkowski and Lavie, 2014) with default parameter settings.

[18] The authors provide no details regarding how this module is combined with other Meteor matchers for alignment and scoring.

arc in the dependency trees of the candidate and the reference. DiscoTK-Party-Tuned (Guzmán et al., 2014), DPMFComb (Yu et al., 2015) and RATATOUILLE (Marie and Apidianaki, 2015) use a learnt combination of the scores from different evaluation metrics. We note that the number and complexity of the metrics used in the above approaches is quite high. For instance, DPMFComb is based on 72 separate evaluation metrics, including the resource-heavy linguistic metrics from the Asiya Toolkit (Giménez and Màrquez, 2010).

As can be seen in Table 4.10, the UPF-Cobalt beats string-level metrics (Group I) by a large margin and achieves a significant improvement over other approaches that make use of linguistic information (Group II). The approach is only outperformed by some of the metrics from Group III, which are learnt combinations of the scores from different evaluation metrics.[19]

4.4.2 Adequacy and Fluency Judgments: MTC-P4 Dataset

Ranking judgments provide little insight regarding how well the magnitude of the differences in quality between the MT outputs of different source sentences is reflected in sentence-level automatic evaluation. To test the metric's performance on absolute quality judgments, we conduct experiments with the MTC-P4 Chinese–English dataset, produced by Linguistic Data Consortium (LDC2006T04). This dataset contains 919 source sentences from the news domain, 4 reference translations and MT outputs generated by 10 translation systems. The translations produced by 6 of the systems were assigned quality scores following the Linguistic Data Consortium evaluation guidelines Linguistic Data Consortium (2005), based on fluency and adequacy criteria, on a 5-point scale. In total, human assessment is provided for 5,514 MT sentences.

Fluency and adequacy scores are normally averaged to obtain global quality scores. We report sentence-level Pearson correlation with the averaged scores, as well as for fluency and adequacy scores separately. We compare the performance of our metric with BLEU (Papineni et al., 2002) and Meteor (Denkowski and Lavie, 2014).

The MTC-P4 dataset contains 4 different human reference translations. The metrics are evaluated in both single-reference and multi-reference scenarios. For the case when only one human reference is used, the reference is chosen at random and is the same for all the evaluation metrics. BLEU was specifically designed to be used with multiple references. It counts the n-gram matches between the MT and any of the available human translations. To adapt Meteor and UPF-Cobalt to the multi-reference scenario, we follow a simple approach of selecting for each sentence the highest of the 4 sentence-level scores obtained with different references (see Qin and Specia (2015) for a description

[19] A probable explanation for the differences in the average correlation achieved by different metrics (with the best performance being achieved on the WMT15 data, and the worst results shown on the WMT13 data) is the difference in inter-annotator agreement in different years.

	Single Reference			Multiple References		
Metric	A	F	Avg	A	F	Avg
UPF-Cobalt	0.460	0.279	0.418	0.491	0.306	0.450
Meteor	0.450	0.262	0.405	0.488	0.302	0.447
BLEU	0.295	0.200	0.278	0.342	0.252	0.332

Table 4.11: Sentence-level evaluation results on MTC4-P4 dataset in terms of Pearson correlation with Adequacy (A), Fluency (F) and Averaged (Avg) adequacy and fluency judgments

of alternative strategies). The results are summarized in Table 4.11.

First, we observe that UPF-Cobalt outperforms BLEU and Meteor for adequacy, fluency and averaged human judgments, in single-reference as well as in multi-reference scenario. The differences between UPF-Cobalt and BLEU were found to be significant in all cases. The differences between UPF-Cobalt and Meteor were found to be significant for fluency scores and average scores in the single-reference scenario.[20]

Secondly, all the metrics present a lower correlation for fluency. The reason is that neither of the reference-based evaluation metrics explicitly addresses this aspect of translation quality. However, BLEU and Meteor are outperformed by UPF-Cobalt in terms of the correlation with fluency judgments. The reason is that syntactic similarity between MT and the reference reflects, although indirectly, the MT fluency. In general, adequacy and fluency are related aspects. If the MT is very similar to a reference, it is probably well-formed. Thus, a metric that is better for predicting adequacy will also show an improvement in predicting fluency judgments.

Finally, the results show that the benefit of using multiple references is much higher in the case of BLEU. This is not surprising, since the evaluation metrics that allow for fuzzy matches between words and constructions are designed precisely to overcome the limitations of using single reference as benchmark. Furthermore, the difference between UPF-Cobalt and Meteor is minimal in the case of multi-reference evaluation. This suggests that the gain in performance achieved by UPF-Cobalt in the single-reference scenario is related to addressing the issue of acceptable variation between the candidate translation and the human reference.

4.4.3 Adequacy Judgments on a Continuous Scale: WMT16 Dataset

We use the so called "direct assessments" data from the WMT16 Metrics Task (Bojar et al., 2016). At WMT16, besides the traditional ranking judgments, absolute quality judgments were collected. The judgments were collected according to the adequacy

[20]The Hotelling-Williams (Williams, 1959) test for dependent correlations was used for significance testing.

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en
UPF-Cobalt	0.652	0.490	0.550	0.616	0.556	0.626
BLEU	$0.568^{\dagger}$	0.447	$0.433^{\dagger}$	$0.499^{\dagger}$	$0.470^{\dagger}$	$0.538^{\dagger}$
Meteor	0.645	0.517	0.540	0.587	0.548	0.618
TER	$-0.578^{\dagger}$	$-0.468^{\dagger}$	$-0.411^{\dagger}$	$-0.441^{\dagger}$	$-0.459^{\dagger}$	$-0.491^{\dagger}$
MPEDA	0.644	0.513	0.538	0.587	0.545	0.616
DPMFcomb	$0.713^{\dagger}$	$0.598^{\dagger}$	$0.584^{\dagger}$	0.627	$0.615^{\dagger}$	$0.663^{\dagger}$
chrF2	0.658	0.469	$0.457^{\dagger}$	$0.581^{\dagger}$	0.534	$0.556^{\dagger}$
BEER	0.661	0.471	$0.462^{\dagger}$	$0.551^{\dagger}$	0.533	$0.545^{\dagger}$

Table 4.12: Sentence-level Pearson correlation with direct assessments from the WMT16 dataset

criterion following the procedure described in Graham et al. (2015) for all available into-English language pairs. Human assessors were asked how much of the meaning of the reference translation was preserved in the MT output. The evaluation was performed using a 0-100 rating scale. Human assessment scores were standardized according to an individual annotator's overall mean and standard deviation. Up to 15 assessments were collected for each MT output from different assessors and the results were averaged to obtain the final score. 560 MT segments sampled randomly from the data were annotated by humans for each language pair, resulting in a total of 3,360 segments of into-English translations. (See Section 2.2.2 for details).

In Table 4.12 we reproduce the results of the task from Bojar et al. (2016) for the best performing systems as well as for the baselines BLEU, Meteor and TER. Metric correlations significantly different from the correlation attained by UPF-Cobalt are marked with †. As before, the significance of the differences in metric performance is computed using the Hotelling-Williams test for the significance of the difference in dependent correlations (Williams, 1959; Graham et al., 2015). Appendix A shows the scatter plots for the scores from the metrics participating in WMT16 Metrics Task for into-English translation directions.

The difference in the results and the significance of the differences between ranking and direct assessments may be due to the small size of the direct assessments dataset, but also to the inherent differences between the two tasks. As discussed in Section2.2.1, ranking and scoring tasks present different challenges for automatic evaluation. If a certain feature of MT output is what frequently distinguishes system translations and this feature is well captured by the metric, such metric will obtain a high correlation for ranking, while this may not necessarily be the case for the scoring task.

4.5 Analysis

In the previous Section, we discussed the performance of UPF-Cobalt in general terms comparing the results with other state-of-the-art-evaluation metrics. However, correlation with human judgments as the only measure of metrics performance may be insufficient to properly understand their advantages and limitations (Amigó et al., 2009). In this Section we provide a further analysis of various aspects of the performance of our metric.

4.5.1 Lexical Resources

To appreciate the influence of different lexical similarity resources on candidate-reference alignment, we computed the number of aligned words corresponding to different types of matches. Table 4.13 shows the number of matches of same word forms (Exact), stem match (Stem)[21], WordNet synsets (Synonym), paraphrase database (Paraphrase), distributional similarity (Distributional) and Named Entities (NE) (see Section 4.3.2), taking into account the number of function (Function) and content (Content) words. The results are computed for the Czech–English translation direction.

First of all, we observe that the total proportion of aligned words is higher in the case of UPF-Cobalt alignment. The data provided in Table 4.13 is descriptive, i.e. an increased number of aligned words does not necessarily imply better alignment quality. Nevertheless, as the use of additional lexical similarity resources was aimed at maximizing lexical coverage, this goal is achieved.

In both cases, the most frequent type of lexical match is the exact match. The percentage of exact matches is higher in the case of Meteor. This difference may be due to the fact that in the case of UPF-Cobalt, function words with no contextual evidence are not aligned. The difference between the proportion of stem matches is due to the fact the in the case of UPF-Cobalt both stem and lemma alignments are considered. There can be no comparison between the difference in the proportion of paraphrase matches, as different paraphrase resources are used and in the case of UPF-Cobalt, only 1-to-1 alignments are allowed. Finally, matches based on distributional similarity represent a considerable percentage of the total number of matches, and as we will see in Section 4.5.2 significantly affect the performance of the metric.

Different types of matches vary in terms of the reliability of the lexical resources, i.e. the confidence that the match between the words is indicative of MT quality. Table 4.14 shows the average context penalty value and the number and percentage of words with context penalty[22] for each type of lexical matches. The highest context penalty is for the words aligned using distributional similarity. Thus, the results confirm our

[21] This corresponds to stem and lemma match in the case of UPF-Cobalt.

[22] The averages are computed with raw context penalty values (see Section 4.3.2).

	UPF-Cobalt		Meteor	
Exact	470,826	77.1 %	480,233	84.4 %
Stem	22,826	3.7 %	7,222	1.3 %
Synonym	16,429	2.7 %	14,592	2.6 %
Paraphrase	25,803	4.2 %	67,117	11.8 %
Distributional	70,629	11.6 %	-	-
NE	4,230	0.7 %	-	-
Total Aligned	610,743		569,164	
Total MT	778,653	78.4 %	778,653	73.1 %
Total Reference	783,516	77.9 %	783,516	72.6 %

Table 4.13: Number of different types of lexical matches found in UPF-Cobalt and Meteor alignment for WMT16 data Czech–English translation direction

intuition, that the noisier the resource, the more probable it is that the aligned words do not have the same function in the candidate and reference sentences, and therefore, may be indicative of errors.[23]

	Average	Count	
Exact	0.249	187,079	39.7 %
Stem	0.612	18,634	81.6 %
Synonym	0.630	13,231	80.5 %
Paraphrase	0.524	18,895	73.2 %
Distributional	0.663	57,790	81.8 %
NE	0.576	3,656	86.4 %

Table 4.14: Average context penalty values for different types of lexical matches for WMT16 data Czech–English translation direction

4.5.2 Ablation Study

In addition to the overall evaluation, we performed a series of ablation tests in order to assess the impact of the individual components of UPF-Cobalt. Each row of Table 4.15 shows the Pearson correlation with human scores when one of the components or features of the metric is excluded. Similarly, Kendall's Tau with WMT16 ranking judgments for the same variants of the metric is reported in Table 4.16[24].

[23] High values for NE alignments can be fixed by introducing a special procedure for counting context penalty only for the head word of the NE. We leave this for future work.

[24] See Tables A.5-A.7 in Appendix A for the results for the WMT14-WMT16 datasets with confidence intervals.

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en
UPF-Cobalt	0.652	0.490	0.550	0.616	0.556	0.626
aligner off	0.653	0.494	0.550	0.600	0.573	$0.606^{\dagger}$
penalty off	0.645	$0.441^{\dagger}$	$0.492^{\dagger}$	0.599	$0.532^{\dagger}$	$0.568^{\dagger}$
weights off	0.652	$0.481^{\dagger}$	0.555	0.615	0.553	$0.623^{\dagger}$
equivalence off	0.652	0.489	0.552	0.616	0.556	0.626
WE off	0.657	0.494	0.544	0.614	$0.581^{\dagger}$	0.622

Table 4.15: Results of ablation test with WMT16 direct assessment judgments

Metric	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	Average
UPF-Cobalt	0.364	0.435	0.388	0.351	0.392	0.382	0.385
aligner off	0.318	0.419	0.362	0.335	0.372	0.338	0.357
penalty off	0.342	0.401	0.394	0.342	0.373	0.374	0.371
weights off	0.357	0.433	0.382	0.355	0.388	0.375	0.382
equivalence off	0.357	0.436	0.389	0.348	0.389	0.382	0.384
WE off	0.349	0.437	0.385	0.359	0.387	0.354	0.378

Table 4.16: Results of ablation test with WMT16 ranking judgments

In the first place, we compared the performance of the metric when using Meteor aligner instead of MWA ("aligner off" in Tables 4.15 and 4.16). This results in a drop in performance for all translation directions in the case of the ranking task and a significant drop in performance for the Turkish–English translation direction in the case of WMT16 direct assessments.

As we suggested in Section 4.1, using word context to avoid aligning spurious matches between function words in candidate and reference translations indeed appears beneficial for MT evaluation. Consider the following example in Table 4.17 as an illustration. Aligned words are marked in bold. The personal pronoun "us" in the reference is not aligned to acronym "US" in MWA alignment, since these two words do not share any aligned words in their syntactic or textual contexts. The same regards the alignment of the copula verb, as the arguments are completely different in the candidate and reference translations.

An advantage of Meteor's aligner is the handling of phrase alignments. It must be noted that we did not adapt context penalty to phrase matches. This could have resulted in a higher correlation for Meteor's aligner. We leave the integration of phrase level alignments in the metrics, as well as the adaptation of context penalty computation for future work, as this functionality is highly relevant for the evaluation task, as it allows covering acceptable variation that involves multi-word expressions.

One of the advantages of UPF-Cobalt is the integration of distributional similarity in monolingual alignment (we will refer to this component as Word Embeddings (WE)

Source:	Bize bir misyon sebebiyle bir süre göründü.
MT_{MWA}:	**A** US **mission** was due **to appear for** a while.
MT_{METEOR}:	**A US mission was** due **to appear** for a while.
Ref:	It seemed to us that he was on a mission.

UPF-Cobalt	0.3031
aligner off	0.4342

Table 4.17: Example of spurious function word matches (WMT16 dataset, Turkish–English translation, sentence 1139)

for brevity). Example in Table 4.18 illustrates the effect of this component on the scores generated by the metrics. This is an example of an explicitation shift in the reference translation, i.e. the use of a word with a more specific meaning, which can be addressed by the use of distributed word representations. Note that although the difference between the two alignments concerns only one word, the difference in the scores produced by UPF-Cobalt is high because the word involved is the predicate of the sentence and have various dependent words in its context, all of whose scores are therefore affected by the "mistranslation" of this word.

Source:	Au fost duşi la spital
$MT_{MWA_{WE}}$:	**They** have **been taken to hospital**
MT_{METEOR}:	**They** have **been** taken **to hospital**
Ref:	They were rushed to the hospital

UPF-Cobalt	0.7082
WE off	0.4622

Table 4.18: Example of candidate-reference alignment using distributional similarity (WMT16 dataset, Romanian–English translation, sentence 1733)

However, the gains on this dataset are not that evident if the word embedding component is eliminated ("WE off" in Tables 4.15 and 4.16). For ranking WMT datasets, removing the WE component implies a considerable decrease in the correlation. Qualitative analysis of the results shows that its main contribution concerns cases of quasi-synonyms, i.e. words that can be considered synonymous only given the similarity of their contexts. The noise introduced by the component is neutralized by context penalty. If unrelated words are aligned, their context penalty will be high and aligning them will not increase sentence-level evaluation score. Also, in the ranking formulation of the evaluation task, distributional similarity helps to discriminate between low-quality translations. That is to say, it allows distinguishing sentences where words are at least

minimally related from sentences, in which, for instance, source-language words are simply left untranslated. For the direct assessment data, the gains from the WE component is not evident. Specifically, for Russian–English translation direction in the direct assessment dataset the performance is actually significantly improved when the WE component is eliminated. As we will see below, the metric is overly permissive with relatively high quality translations, when it is unable to properly penalize lexical or word order differences.

Eliminating the component of dependency equivalence mapping ("equivalence off" in Tables 4.15 and 4.16), i.e. treating all mismatches between dependency labels as context differences, produces a smaller decrease in the correlation. Representing syntactic context as immediate neighbors of the word in a dependency graph allows covering a limited set of equivalent constructions, which are not frequent enough to have a significant impact on the results (see Table 4.19). The framework is flexible and more complex context equivalence definitions can be integrated in the future. Also, this means that the metric can be used independently of the language pair given the availability of syntactic parsing.

	Changed	Total	%
cs-en	6,806	35,988	19
de-en	5,768	29,990	19
fi-en	5,585	27,000	21
ro-en	3,123	13,993	22
ru-en	7,751	29,980	26
tr-en	4,464	24,000	19

Table 4.19: Number of sentence scores changed by using equivalent dependency relations in WMT16 direct assessment dataset

Addressing acceptable variation between translation variants is naturally limited. When human translation is substantially different from the source sentence, semantic equivalence between the candidate and reference translations can hardly be established. In the next Chapter we suggest a method for addressing this issue.

To test if giving different weights to contextual differences according to the dependency functions of the words involved, we put the values of all the weights to 1 ("weights off" in Tables 4.15 and 4.16). This negatively affects the results, confirming that some differences are stronger indicators of MT errors than others. Thus, using the proposed weighting scheme, the metric is capable of discriminating more or less serious MT errors based on the relative importance of mistranslated material.

Finally, we eliminated the context penalty component ("penalty off") in Tables 4.15 and 4.16, thus the final score is an F-measure over the number of lexical matches of different types. Evidently, this results in a significant drop in the correlation. Note that such a strategy achieves very high correlation for the task of sentence similarity (Aker

et al., 2016). However, MT evaluation has a crucial difference consisting in the presence of different types and combinations of translation errors which impact the perceived translation quality to a varying extent.

4.5.3 Language Pairs

One of the question that arose during the analysis of the results concerns the variation in the metric performance for different language pairs. Note that although we work with into-English translation direction, and thus the target language is always the same, the source languages are different, and therefore, MT errors and the overall translation quality is different as well. First of all, we note that the performance of all the metrics varies depending on the language pair, with some translation directions having higher correlation on average than others. An inspection of the results for different years of WMT Metrics Tasks shows that one of the reasons actually does not have anything to do with the metrics but with the reliability of human evaluation. Table 4.20 provides inter-annotator agreement between metric and human rankings as well as the same measure for human rankings of different annotators.[25] As can be seen in Table 4.20, the results are highly correlated. As a matter of fact, the level of agreement between annotators can be treated as an upper bound for the performance of evaluation metrics.

LP	$\text{Human}^{\text{TIES}}$	$\text{Human}^{-\text{TIES}}$	Metric
cs-en	0.458	0.688	0.495
de-en	0.423	0.667	0.482
fi-en	0.388	0.628	0.445
fr-en	0.343	0.609	0.395
ru-en	0.372	0.659	0.418

Table 4.20: Comparison of inter-annotator agreement between human annotators and evaluation metrics for the WMT15 dataset

Thus, the absolute difference between the metrics' correlation for different language pairs is related to the differences in the inter-annotator agreement. On average, independently of the metric the correlation is always higher for the language pairs with higher inter-annotator agreement. To eliminate the impact of this factor, we examined the behavior of UPF-Cobalt in relative terms, testing the difference in the correlation with human judgments against the baseline BLEU and the best of the participating metrics for

[25] In the WMT formulation, the Kappa measure of inter-annotator agreement is equivalent to the Kendall's Tau correlation (see Section 2.2.3).

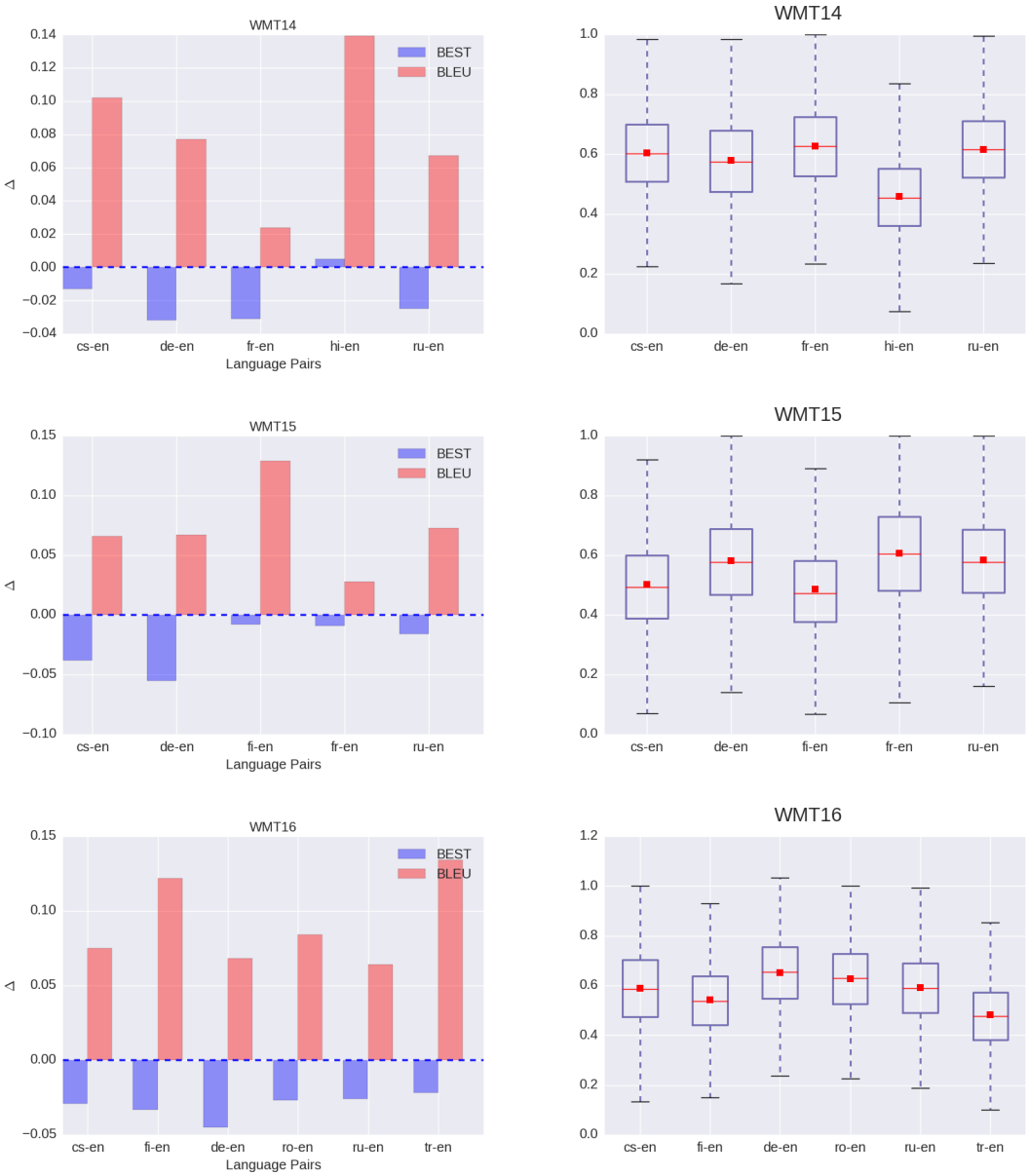


Figure 4.3: Left: Difference in Kendall's Tau correlation with BLEU and with the best performing metric. Right: Distribution of UPF-Cobalt scores across different language pairs

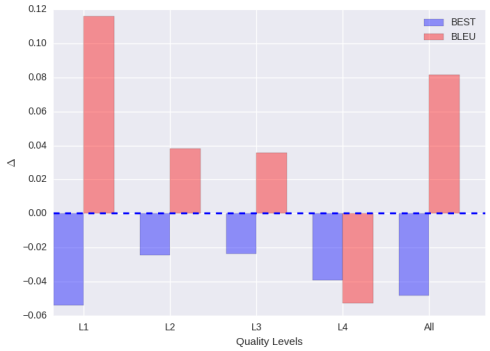


Figure 4.4: Difference in performance between UPF-Cobalt and benchmark metrics on data samples with different levels of MT quality

each year. The results are shown in Figure 4.3[26] suggest that independently of the particular language pair, the metric performs better for the data with lower average scores (hi-en in WMT14, fi-en in WMT15 and tr-en in WMT16). To confirm this observation we analyzed the performance of the metrics on the data samples with different levels of underlying MT quality.

4.5.4 Quality Levels

In order to inspect the performance of our metric for different levels of MT quality, we sort and split the WMT16 direct assessment dataset using 4 quantiles based on human scores. This results in four groups: top 25%, mid-high 25%, mid-low 25% and bottom 25% of the data (which we call L1-L4), each containing 840 sentences. For each group, we measure Pearson correlation between metric and human scores.

Figure 4.4 shows the difference in performance between UPF-Cobalt and benchmark metrics on data samples with different levels of MT quality (BEST stands for the best performing metric on the WMT16 dataset, DPMFcomb). UPF-Cobalt attains no gains in performance for high-quality MT outputs. The reason may be twofold. On the one hand, surface-level metrics seem to perform relatively well on high quality translations. On the other hand, UPF-Cobalt may be excessively permissive with translations of moderate quality.

To test this, we did the same ablation procedure as in Section 4.5.2, and computed correlation for those different quality levels. Table 4.21 presents the results. For high quality translation a significant increase in correlation can be observed when distribu-

[26] All the differences in the means except for the cs-en vs. ru-en language pairs in the WMT16 dataset were found to be statistically significant with the T-test for the means of two independent samples.

tional similarity and syntactic equivalence components are eliminated ("WE off" and "equivalence of" in Table 4.21). These components are, therefore, less suitable to discriminate between high quality translations. A careful inspection of the data shows that the reason is that the metric is overly permissive with the differences in lexical choice and word forms. Consider the following example from Czech–English translation from the WMT16 dataset (Table 4.22). MT and $MT_{\text{WE off}}$ show the output of an MT system with the aligned words marked in bold for the full UPF-Cobalt metric and for UPF-Cobalt with the distributional similarity component eliminated, respectively. The Table below shows the corresponding evaluation scores for the full metric (UPF-Cobalt), ablation of the distributional similarity component (WE off) and ablation of equivalent dependency types (equivalence off), as well as the corresponding human direct assessment scores (DA). Comparing raw scores may be inaccurate, since the scales and distributions of the scores are different. Therefore, in addition to the raw scores, we provide standard scores (z-score) and ranks (rank). The ranks are obtained by sorting the MT outputs from best to worst according to the corresponding set of scores.

	L1	L2	L3	L4
UPF-Cobalt	0.146	0.121	0.160	0.406
meteor aligner	0.171	0.147	$0.131^{\dagger}$	0.423
penalty off	$0.210^{\dagger}$	$0.097^{\dagger}$	0.153	$0.346^{\dagger}$
weights off	0.141	0.124	0.154	0.403
equivalence off	0.143	$0.115^{\dagger}$	0.159	$0.415^{\dagger}$
WE off	0.149	0.116	0.158	$0.421^{\dagger}$

Table 4.21: Ablation test for MT outputs of different quality on WMT16 direct assessment data

The equivalence between the phrases "village symbol" in the reference translation and "symbol of community" in the candidate translation is established using word embeddings to match the words "village" and "community". Equivalent dependency types are used to recognize the equivalence between the relations of preposition "of" and noun compounding between the words "symbol" and "community", and "symbol" and "village" in the candidate and reference translations respectively. Without this component, the words involved receive a high context penalty. The difference between the MT output and human translation, however, is penalized in manual evaluation.[27]

Along the same lines, some of the disagreement between UPF-Cobalt and direct

[27]Note that the reference translation contains a shift in the verb tense, resulting in the so called historical present tense, often used in news headlines in English. On the contrary, the original past tense is preserved in the MT output. Evidently, the fact that the sentence is a headline is unknown to the MT system. This illustrates the point made in the previous Chapter regarding candidate-reference differences that are not due to MT errors but are an artifact of the characteristics of human translation inaccessible to current MT systems.

Source:	Symbol obce ožil.
MT:	A **symbol** of **community came to life**.
$MT_{WE\ off}$:	A **symbol** of community **came to life**.
Ref:	Village symbol comes to life

	score	z-score	rank
UPF-Cobalt	0.7951	0.5203	215
WE off	0.6531	-0.2035	433
equivalence off	0.6982	-0.0847	395
DA	0.4162	-1.398	832

Table 4.22: Example of the effect of the WE component and the syntactic equivalence component on the performance of UPF-Cobalt (WMT16 dataset, Czech–English translation, sentence 1525)

assessment scores comes from ignoring word form and word order differences that, as a matter of fact, affect human scores, although they do not affect the syntactic parse, as illustrated in Table 4.23. In this case, a subject-predicate agreement error in the MT output ("was" instead of "were") has a very small effect on UPF-Cobalt score, as the syntactic parse is not affected by this error, and the equivalence between passive/active alternation still holds and is recognized by the metric. Furthermore, the words "markets" and "venues" are aligned using the distributional similarity component, although they are probably not perceived as equivalent in manual evaluation.

Source:	Solide Gewinne verbuchten auch die Handelsplätze in London und Paris.
MT:	Solid profits was also recorded by the trading venues in London and Paris.
Ref:	The markets in London and Paris also registered solid gains.

	score	z-score	rank
UPF-Cobalt	0.8301	0.7491	147
WE off	0.6924	0.0337	342
equivalence off	0.7488	0.2459	289
DA	0.5165	-0.9402	671

Table 4.23: Example of word form error with a different impact on metric and human scores (WMT16 dataset, German–English translation, sentence 1831)

Similarly, some of the word order differences that affect the perception of the MT quality do not affect the syntactic parse and, therefore, are ignored by the metric, as

shown in the example in Table 4.24. Interestingly, there is a significant improvement in correlation for low quality data when the context penalty component is eliminated. The reason for that may be that a syntactic penalty is less reliable for low quality outputs. However, eliminating the fragmentation penalty from Meteor also results in an improvement for low quality translation. This seems to suggest that when discriminating between low quality translations lexical differences are more important (or reliable) than differences in word order or syntactic structure.

Finally, while in general the correlation is low because making fine-grained distinctions between translations of similar quality is a much more challenging task, we observe that the correlation is generally higher for high quality MT. This means that the metrics discriminate better between high quality translations. A probable explanation is that reference-based systems in general are more reliable when a good match with the reference translation can be found. High quality outputs tend to contain a higher number of matches with the reference, and thus metrics naturally have more information to measure translation quality. By contrast, low quality MT outputs contain very few matches and thus metric scores simply indicate that the MT output is different from the available reference. Another possible reason is that human evaluation may also be more consistent for high quality translation. Low quality MT outputs contain numerous errors which may be perceived in different ways by the annotators, leading to lower inter-annotator agreement and to less reliable scores. This, however, needs to be tested further, which we leave as future work.

Source:	“Альбомы все еще имеют значение”, - сказал он.
MT:	“Albums still matter,” said he.
Ref :	“Albums still matter,” he said.

	score	z-score	rank
UPF-Cobalt	0.8301	0.7491	147
DA	0.5165	-0.9402	671

Table 4.24: Example of word order error with a different impact on metric and human scores (WMT16 dataset, Russian-English translation, sentence 2622)

4.6 Summary

In this Chapter we proposed a new metric for MT evaluation, UPF-Cobalt. The metric leverages lexical similarity and syntactic context to assess the impact of the differences between candidate and reference words on sentence-level MT quality as perceived in manual evaluation.

We used contextually informed alignment for the purpose of MT evaluation, which reduced the number of spurious matches between candidate and reference translations, resulting in a more accurate automatic evaluation.

Inspired by the findings from the previous Chapter, we increase the coverage of acceptable differences at lexical level by using distributed representations of words and at syntactic level by matching equivalent syntactic constructions.

We modeled the effect of candidate-reference differences on perceived translation quality taking various measurements on the pairs of aligned words: (a) number of related matching words, (b) number of syntactic dependents, and (c) type of syntactic function.

We performed an in-depth meta-evaluation study testing the metric's performance in different settings including different MT systems, language pairs, types of human judgments and levels of translation quality. UPF-Cobalt outperforms most of the current approaches at the ranking task. The results for the scoring task are mixed. Our analysis revealed that the metric is overly permissive with moderate quality MT. This could be addressed by including string-level information in the computation of the score. Furthermore, we observed that, as any other reference-based evaluation metric, UPF-Cobalt lacks information regarding the characteristics of non-aligned words. We address these latter limitations in the next Chapter.

Chapter 5

INTEGRATING TRANSLATION FLUENCY INTO AUTOMATIC MT EVALUATION

Reference-based MT evaluation metrics measure translation quality based on the similarity between the MT output and one or various human reference translations. This strategy relies on the assumption that the correct translations of the same sentence are similar to each other. As was very well illustrated in Lommel (2016), on average there are more string-level differences between alternative human translations of the same source sentence than between an MT output and a human translation.

In Chapter 3 we have discussed the sources of variation in human translation in the light of reference-based MT evaluation. We have concluded that, candidate-reference differences may arise not only due to MT errors, but also due to the presence of translation shifts in the reference and a close translation choices in the MT output resulting in unjustifiably low automatic evaluation scores.

A major challenge for reference-based metrics is, therefore, to determine to what extent candidate-reference differences are predictive of MT quality. In the previous Chapter, we described a method for quantifying the contribution of matching words to the sentence-level metric score based on the differences found in their syntactic contexts. Although our approach achieved a significant improvement over various state-of-the-art evaluation metrics (see Section 4.4), not unlike other reference-based metrics, it still lacks information regarding the quality of MT fragments that do not match the reference. In fact, if no reasonable match with the reference can be found, the scores generated by reference-based metrics become meaningless.

In this Chapter we describe our strategy to compensate for this lack of information. To that end, besides computing candidate-reference similarity, we propose to explicitly measure the fluency of MT output. The fluency of the MT fragments that are not aligned to any words in the reference can give us clues regarding the quality of those fragments.

It is more probable that the differences between a fluent MT output and the reference are due to acceptable variation. Conversely, if the MT fragments absent from the reference are ill-formed, it is more probable that the differences are due to MT errors.[1]

As discussed in previous chapters, the core aspects of translation quality are fidelity to the original text (or adequacy, in MT parlance) and acceptability (also termed fluency) regarding the target language norms and conventions (Toury, 2012). In our view, the fluency aspect of translation quality has been overlooked in reference-based MT evaluation. Reference-based metrics are largely focused on MT adequacy, as they do not evaluate the appropriateness of the translation in the context of the target language. Translation fluency is assessed only indirectly, through the comparison with the reference. However, the difference from a particular human translation does not imply that the MT output is disfluent.

Predicting MT quality with respect to the target language norms has been investigated in the field of quality estimation, a different evaluation scenario, when human translations are not available as benchmark (see Section 2.1.4). Along with the features based on the target LM probability of the MT output, which have been widely used for quality estimation (Specia et al., 2009), we designed a more detailed representation of MT fluency that takes into account the number of disfluent segments observed in the candidate translation. To include the fluency aspect in reference-based MT evaluation, we integrated a set of features representing translation fluency with our previous approach described in Chapter 4. We tested our approach with the data available from the WMT Metrics Tasks and obtained very promising results, which rival the best-performing metric submissions.

5.1 CobaltF: A Fluent Metric for Machine Translation Evaluation

We learn an evaluation metric that combines a series of adequacy-oriented features extracted from our reference-based metric UPF-Cobalt described in the previous Chapter with various features intended to focus on translation fluency. This section first describes the metric-based features and then the selection and design of our fluency-oriented features.

[1] We are aware that, especially in the case of data-driven MT, systems can produce perfectly acceptable output which does not preserve the original meaning. However, as will be shown in Section 5.3, when used together with adequacy features, the fluency aspect provides complementary information that improves evaluation accuracy.

5.1.1 Adequacy-oriented Features

Recall from the previous Chapter that UPF-Cobalt integrates in a single score various distinct characteristics of MT output (lexical choice, word order, grammar issues, such as wrong word forms or wrong choice of function words, etc.). We note that these components can be related, to a certain extent, to the aspects of translation quality being discussed in this Chapter. The syntactic context penalty in UPF-Cobalt depends on the well-formedness of MT output, and may reflect, although indirectly, grammaticality and fluency, whereas the proportion of aligned words depends on correct lexical choice (see Section 4.4.2 for the discussion of the metric's performance with respect to these different aspects of translation quality).

Using the components of the metric instead of the scores yields a more fine-grained representation of the relation between MT output and the reference translation. For example, this idea has been explored for the BLEU metric showing that deconstructing its scores into various components results in better learning (Song et al., 2013). We explore this idea in our experiments by designing a decomposed version of UPF-Cobalt. We designed various lexical features that count the number of lexical matches of different types. Also, given that words with high context penalty are assumed to be indicative of translation errors, we extracted various different measurements from word-level context penalty. More specifically, we use 48 features grouped below (see Appendix B):

- Percentage and number of aligned words in the candidate and reference translations
- Percentage and number of aligned words with different levels of lexical similarity in the candidate and reference translations
- Percentage and number of aligned function and content words in the candidate and reference translations
- Minimum, maximum and average context penalty[2]
- Percentage and number of words with high context penalty[3]
- Number of words in the candidate and reference translations

5.1.2 Fluency-oriented Features

The translation process is conditioned by the tension between two prototypical expectations: that of maximal similarity between source and translated texts and that of naturalness of the translated text in the target language. The main requirements that define translation quality are, therefore, on the one hand, its fidelity to the original and, on the

[2]Since here we use context penalty as a feature on its own, we do not need it to be in the range [0-1] and therefore, we use the CP score from Section 4.3.2 without normalization.

[3]These are the words with the context penalty value higher than the average computed on the dataset we used for training (Section 5.2), which approximately equals to 1.0.

other hand, its acceptability with regards to the linguistic norms of the target language (Toury, 2012).

We suggest that the fluency aspect of translation quality has been overlooked in reference-based MT evaluation. Even though syntactically-informed metrics capture structural differences and are, therefore, assumed to account for grammatical errors, we note that the distinction between adequacy and fluency is not limited to grammatical issues and thus exists at all linguistic levels. Fluency determines how well a translated text conforms to the linguistic regularities of the target language, involving both grammatical correctness and the use of stylistically appropriate linguistic constructions. For instance, at lexical level, the choice of a particular word or expression may be similar in meaning to the one present in the reference, but awkward or even erroneous if considered in the context of the norms of the target language use. Conversely, due to the variability of linguistic expression, neither lexical nor syntactic differences from a particular human translation imply ill-formedness of the MT output.

Both in the field of quality estimation, which aims to evaluate MT quality without a reference translation, and in the context of MT development, sentence fluency is commonly described in terms of the frequencies of word n-grams with respect to a target LM. An LM aims to estimate how likely a sequence of words is in a given language. Language Models (LMs) are widely used in many NLP tasks. Along with the translation model, LMs are an essential component of statistical MT systems. Formally, an n-gram LM is a probability distribution $p(W)$ over sequences of words $W = w_1, w_2, ..., w_l$. This distribution is estimated from a large amount of texts counting how often W occurs. Most of long sequences of words, however, will not occur in the data at all or will occur with very low frequency. To deal with the problem of data sparseness, the most widely-used method, n-gram LM, breaks up the process of predicting a word sequence W into predicting one word at a time. The LM probability $p(w_1, w_2, ..., w_l)$ is a product of word probabilities given a history of the preceding words. A crucial assumption, known as Markov assumption that makes n-gram language modeling feasible is that only a limited number of previous words affect the probability of the next word. Thus, the history of the preceding words can be limited to n words, known as the order of the model. The order is chosen depending on how much training data is available, with $n = 3$ being the most common choice. Word probabilities are estimated from the data using maximum likelihood estimation:

$$p(w_i|w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i})}{\sum_{wi} c(w_{i-n+1}^{i})} \tag{5.1}$$

where $c(w_{i-n+1}^{i})$ denotes the number of times that the n-gram w_{i-n+1}^{i} occurs in the data. Then, the probability of a word sequence $p(W)$ can be computed as follows:

$$p(W) = \prod_{i=1}^{l+1} p(w_i | w_{i-n+1}^{i-1}) \tag{5.2}$$

Smoothing techniques are used in order to avoid assigning zero probability to the n-grams that were not encountered in the training data (Chen and Goodman, 1996) (more recently, continuous LMs (Schwenk, 2007)).

An LM is typically evaluated measuring the probability it assigns to a test set. The most common evaluation metric for the LMs is perplexity. Perplexity (PP) is based on cross-entropy ($H(p_{LM})$), defined as:

$$H(p_{LM}) = -\frac{1}{n} \sum_{i=1}^{n} \log p_{LM}(wi | w1, ..., w_{i-1}) \tag{5.3}$$

$$PP = 2^{H(p_{LM})} \tag{5.4}$$

As mentioned above, LM probability and the derived measures, cross-entropy and perplexity, are also often employed in quality estimation as features for predicting word-level and sentence-level quality of the MT output under the assumption that infrequent combinations of words can indicate translation errors. LM probability estimates have been shown to achieve a competitive performance both for pinpointing word-level translation errors (Raybaud et al., 2011) and for predicting sentence-level translation quality (Shah et al., 2013). However, the same measures are used as a component in statistical MT systems, resulting in a circularity in MT development and evaluation. In this work, in addition to the LM-based features that have been shown to perform well for sentence-level quality estimation (Shah et al., 2013), we introduce more complex features derived from word-level n-gram statistics.

Furthermore, besides the word-based representation, we rely on POS tags. As suggested by Felice and Specia (2012), morphosyntactic information can be a good indicator of ill-formedness in MT outputs. As with words, infrequent combinations of POS-tags can be assumed to indicate translation errors. Note that POS-tags result in a less sparse representation than words, potentially providing better generalization.

First, we select 16 simple sentence-level features from previous work (Felice and Specia, 2012; Specia et al., 2010b), summarized below.

- Number of words in the candidate translation
- LM probability and perplexity of the candidate translation
- LM probability of the candidate translation with respect to an LM trained on a corpus of POS tags of words
- Percentage and number of content and function words
- Percentage and number of verbs, nouns and adjectives

Essentially, these features average LM probabilities of the words to obtain a sentence-level measurement. While being indeed predictive of sentence-level translation fluency, they are not representative of the number and scale of the disfluent fragments contained in the MT output. Moreover, if an ill-formed translation contains various word combinations that have very high probability according to the LM, the overall sentence-level LM score may be misleading.

To overcome the above limitations, we use word-level n-gram frequency measurements and design various features to extend them to the sentence level in a more informative way. We rely on LM backoff behaviour, as defined in Raybaud et al. (2011). LM backoff behaviour is a score assigned to the word according to how many times the target LM had to back-off in order to assign a probability to the word sequence. The intuition behind is that an n-gram not found in the LM can indicate a translation error. Specifically, the backoff behaviour value $b(w_i)$ for a word w_i in position i of a sentence is defined as:

$$b(w_i) = \begin{cases} 7, & \text{if } w_{i-2}, w_{i-1}, w_i \text{ exists in the model} \\ 6, & \text{if } w_{i-2}, w_{i-1} \text{and } w_{i-1}, w_i \text{ both exist} \\ & \text{in the model} \\ 5, & \text{if only } w_{i-1}, w_i \text{ exists in the model} \\ 4, & \text{if only } w_{i-2}, w_{i-1} \text{and } w_i \text{ exist} \\ & \text{separately in the model} \\ 3, & \text{if } w_{i-1} \text{and } w_i \text{ both exist} \\ & \text{in the model} \\ 2, & \text{if only } w_i \text{ exists in the model} \\ 1, & \text{if } w_i \text{ is an out-of-vocabulary word} \end{cases} \tag{5.5}$$

We compute this score for each word in the MT output and then use the mean, median, mode, minimum and maximum of the backoff behaviour values as separate sentence-level features. Also, we calculate the percentage and number of words with low backoff behaviour values (< 5) to approximate the number of fluency errors in the MT output.

Furthermore, we introduce a separate feature that counts the words with a backoff behaviour value of 1, i.e. the number of out-of-vocabulary (OOV) words (words that were not found in the LM). OOV words may be indicative of the cases when source words are left untranslated in the MT output. Consider the example in Table 5.1. Neither of the candidate translations contains a good match with the reference. However, it is clear that MT1 output is more acceptable than MT2. Whereas the former contains an expression that shares some aspects of meaning with the reference translation, the latter contains an untranslated word. Our intuition is that OOVs have a high negative impact on the human perception of MT quality. This still happens quite a lot in the

current MT systems. Table 5.2 provides the average counts of OOVs per sentence in the WMT14-WMT16 datasets. (See Section 5.2 for the details regarding the LM used in our experiments).

Source:	Sešly jsme se.
MT1	We came together.
MT2	We sešly.
Ref:	We met.

Table 5.1: Example of an out-of-vocabulary word from the WMT2014 Czech–English dataset, sentence 2467)

Evidently, the fact that a word is not found in the LM and thus is considered OOV does not necessarily mean that this is a word in the foreign (source) language. This can also be a proper name or simply a very infrequent word absent from the data which was used for training the LM.

Finally, as noted before, UPF-Cobalt, not unlike the majority of reference-based metrics, lacks information regarding the MT words that are not aligned or matched to any reference word. Such fragments do not necessarily constitute an MT error, but may be due to acceptable linguistic variations. Collecting fluency information specifically for these fragments may help to distinguish acceptable variations from MT errors. If a candidate word or phrase is absent from the reference but is fluent in the target language, then the difference is possibly not indicative of an error and should be penalized less. Based on this observation, we introduce a separate set of features that compute the word-level measurements discussed above only for the words that are not aligned to the reference translation. Consider how this affects the issue of proper nouns counted as OOV mentioned above. If a word is not aligned to the reference translation and not found in the LM, we can be more confident that it is a non-translated word indicative of an MT error. If on the other hand, the word is not found in the LM because it is, for instance, a proper noun, in the case of a correct translation, it will probably be aligned to the reference.

This results in 49 additional features grouped below (see Appendix B):

- Summary statistics of the LM backoff behaviour (word and POS-tag LM)
- Summary statistics of the LM backoff behaviour for non-aligned words only (word and POS tag LM)
- Percentage and number of words with low backoff behaviour value (word and POS tag LM)
- Percentage and number of non-aligned words with low backoff behaviour value (word and POS tag LM)
- Percentage and number of OOV words

	WMT14	WMT15	WMT16
cs-en	3.939	3.751	0.918
de-en	3.891	3.841	1.133
fi-en	-	4.246	1.225
fr-en	4.17	3.11	-
hi-en	4.361	-	-
ro-en			1.164
ru-en	3.889	3.826	1.056
tr-en	-	-	1.511

Table 5.2: Average number of out-of-vocabulary words per sentence in WMT14-WMT16 into-English datasets

- Percentage and number of non-aligned OOV words

5.2 Experimental Setup

Our model is a simple linear interpolation of the features presented in the previous sections. The score for a candidate translation $score(cand)$ is computed as:

$$score(cand) = \vec{w} \cdot \vec{x}_{cand} \tag{5.6}$$

where $\vec{w}$ is a weight vector and $\vec{x}_{cand}$ is a vector of feature values for the candidate translation $cand$. For training and testing, we use the data available from the WMT14-WMT16 Metrics Tasks (see Section 2.2.2). To be able to train on ranking data and produce absolute scores at test time, we use the learn-to-rank approach (Burges et al., 2005) for tuning the weights. This strategy was used in Hopkins and May (2011) for statistical MT parameter tuning and later successfully applied for MT evaluation (Guzmán et al., 2014; Stanojevic and Sima'an, 2015).

Given a pairwise comparison between two candidate translations where one translation ($cand_{good}$) is judged to be better than the other ($cand_{bad}$), a successful MT evaluation metric scores must be $score(cand_{good}) > score(cand_{bad})$. Therefore:

$$\begin{aligned} score(cand_{good}) > score(cand_{bad}) &\Leftrightarrow \\ \vec{w} \cdot \vec{x}_{good} > \vec{w} \cdot \vec{x}_{bad} &\Leftrightarrow \\ \vec{w} \cdot \vec{x}_{good} - \vec{w} \cdot \vec{x}_{bad} > 0 &\Leftrightarrow \\ \vec{w} \cdot (\vec{x}_{good} - \vec{x}_{bad}) > 0 & \\ \vec{w} \cdot (\vec{x}_{bad} - \vec{x}_{good}) < 0 & \end{aligned} \tag{5.7}$$

The learning task can be then formulated as a binary classification problem, the two feature vectors $\vec{w} \cdot (\vec{x}_{good} - \vec{x}_{bad}) > 0$ and $\vec{w} \cdot (\vec{x}_{bad} - \vec{x}_{good}) < 0$ being considered as

positive and negative training instances, respectively.

We use a standard implementation of the Logistic Regression algorithm from the Python toolkit `scikit-learn`[4]. For the extraction of word-level backoff behaviour values and sentence-level fluency features, we use `Quest++`[5], an open source tool for quality estimation (Specia et al., 2015). We employ the LM used to build the baseline system for the WMT15 Quality Estimation Task (Bojar et al., 2015).[6] This LM was trained on the data from the WMT12 Translation Task (a combination of News and Europarl data) and thus matches the domain of the datasets used in our experiments. POS tagging was performed with TreeTagger (Schmid, 1999).

We experimented with combining adequacy-oriented and fluency-oriented features (Cobalt-F system) and combining the scores of the metrics BLEU, Meteor and UPF-Cobalt with fluency-oriented features (Metrics-F system). The results are presented below.

5.3 Experimental Results

We participated in the WMT16 Metrics Task (Bojar et al., 2016). In addition to the ranking setting from the previous years, direct assessment scores based on the adequacy criterion were collected following the procedure described in Graham et al. (2015) for all available into-English language pairs. See Section 2.2.2, for a detailed description of this dataset.

In Table 5.3 we reproduce the results of the task from Bojar et al. (2016) for the best performing evaluation metrics, as well as for the standard BLEU and Meteor metrics. Specifically, Table 5.3 shows the correlation of sentence-level metric scores with two human evaluation variants: τ stands for Kendall's τ computed over WMT relative rankings (RR); and r stands for Pearson correlation coefficient computed between metric scores and direct assessments of absolute translation adequacy (DA).[7] Metric correlations for the direct assessments not significantly outperformed by any other metric are highlighted in bold. As before, the significance of the differences in metric performance is computed using the Hotelling-Williams test for the significance of the difference in dependent correlations (Williams, 1959; Graham et al., 2015). The metrics participating in the task were described in Section 4.4.1. Here we note that our approach attains the best performance on the scoring task, sharing the first place with DPMFComb evaluation system. The results of the ranking task need further investigation, since no reliable

[4] `http://scikit-learn.org/`

[5] `https://github.com/ghpaetzold/questplusplus`

[6] `http://www.statmt.org/wmt15/quality-estimation-task.html`.

[7] In the metrics task results paper by Bojar et al. (2016), in the metric correlation results for into-English translation the data for German–English and Finnish–English translation directions are placed in the opposite way.

method for verifying confidence intervals for WMT formulation of the ranking task has yet been proposed (Bojar et al., 2016) (The results of WMT ranking tasks are reported with confidence intervals in Appendix A, but those should be taken as tentative).

Direction	cs-en		fi-en		de-en		ro-en		ru-en		tr-en	
Human Gold	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA
# Assessments	70k	12k	15k	12k	19k	14k	11k	12k	18k	13k	7k	13k
# Translations	8.6k	560	2.4k	560	4.6k	560	2.2k	560	4.7k	560	2.2k	560
Correlation	τ	r	τ	r	τ	r	τ	r	τ	r	τ	r
DPMFComb	.388	**.713**	.420	**.584**	.481	**.598**	.383	.627	.420	**.615**	.401	**.663**
Metrics-F	.345	**.696**	.421	**.601**	.447	.557	.388	**.662**	.412	**.618**	.424	**.649**
Cobalt-F	.336	.671	.415	**.591**	.433	.554	.361	.639	.397	**.618**	.423	.627
UPF-Cobalt	.359	.652	.387	.550	.436	.490	.356	.616	.394	.556	.379	.626
BEER	.342	.661	.371	.462	.416	.471	.331	.551	.376	.533	.372	.545
MPEDA	.331	.644	.375	.538	.425	.513	.339	.587	.387	.545	.335	.616
ChrF2	.341	.658	.358	.457	.418	.469	.344	.581	.383	.534	.346	.556
UOW-REVAL	.261	.577	.329	.528	.376	.471	.313	.547	.314	.528	.342	.531
BLEU	.284	.557	.265	.448	.368	.484	.272	.499	.330	.502	.245	.532
Meteor	.329	.645	.379	.540	.426	.517	.341	.587	.385	.548	.342	.618

Table 5.3: Sentence-level evaluation results from WMT16 Metrics Tasks for into-English language pairs

To appreciate the advantages of the use of fluency features, consider the example in Table 5.4. According to human rankings from WMT16 dataset, MT1 is better than MT2. The latter is closer to the reference translation in terms of string-level similarity. However, it contains a fluency error resulting from a mistranslation of Czech impersonal verb "jednat se" ("to be about something"). The former contains an acceptable variation with respect to the reference translation: "Either way" - "In any case". UPF-Cobalt does not recognize acceptable variation and assigns a lower score to MT1, while the fluency error is not penalized appropriately. The integration of fluency features results beneficial here, as the MT1 fragment that is not found in the reference is fluent, whereas MT2 contains an improbable sequence of function words "the is a" (or the corresponding POS-tags). As a result, a higher score is correctly assigned to MT1 by Cobalt-F.

5.3.1 Further Analysis

We conduct a further analysis of our feature-based approach using the data from the WMT14-WMT15 Metrics Tasks.

As in the previous experiments, for evaluation we use the Kendall correlation coefficient (τ) with human preference judgments (see Section 4.4.1). Table 5.5 summarizes the results. Group I presents the results achieved by UPF-Cobalt and its decomposed version described in Section 5.1.1. Contrary to our expectations, the performance is

Source:	Každopádně se jedná o skandální postup.
Ref:	It was, in any case, a scandalous procedure.
MT1:	Either way, it is a scandalous procedure.
MT2:	In any case the is a scandalous procedure.

	UPF-Cobalt		Cobalt-F	
	score	z-score	score	z-score
MT1	0.6616	0.4247	2.2467	1.3950
MT2	0.7887	1.1694	1.7306	0.9086

Table 5.4: Example of candidate and reference translations with the corresponding UPF-Cobalt and Cobalt-F scores illustrating the contribution of fluency features (WMT16 dataset, Czech–English translation, sentence 294)

	Metric	cs-en	de-en	fi-en	fr-en	ru-en	Avg τ
I	UPF-Cobalt	.457	.427	.437	.386	.402	.422
	UPF-Cobalt$_{comp}$	.442	.418	.428	.387	.388	.413
II	Features-F	.373	.337	.359	.267	.263	.320
	Cobalt-F$_{simple}$	.487	.445	.455	.401	.395	.437
	Cobalt-F$_{comp}$	.481	.438	.464	.403	.395	.436
	Metrics-F	.502	.457	.450	.413	.410	**.447**
	Metrics	.482	.457	.441	.403	.410	.439
III	DPMFcomb	.495	.482	.445	.395	.418	.447
	BEER_Treepel	.471	.447	.438	.389	.403	.429
	RATATOUILLE	.472	.441	.421	.398	.393	.425
IV	BLEU	.391	.360	.308	.358	.329	.349
	Meteor	.439	.422	.406	.380	.386	.407

Table 5.5: Sentence-level evaluation results for WMT15 dataset in terms of Kendall rank correlation coefficient (τ)

slightly degraded when using the metrics' components (UPF-Cobalt$_{comp}$). Our intuition is that this happens due to the sparseness of the features based on the counts of different types of lexical matches.

Group II reports the performance of the fluency features presented in Section 5.1.2. First of all, we note that these features on their own (Features-F) achieve a reasonable correlation with human judgments, showing that fluency information is often sufficient to compare the quality of two candidate translations. Secondly, fluency features yield a significant improvement when used together with the metrics' score (Cobalt-F$_{simple}$) or with the components of the metric (Cobalt-F$_{comp}$). We further boost the performance by combining the scores of the metrics BLEU, Meteor and UPF-Cobalt with our fluency features (Metrics-F). For the sake of comparison, we present the results for the combi-

nation of the scores from the three metrics without fluency features (Metrics). Note that the results for Metrics are considerably higher than UPF-Cobalt. Recall from Section 4.5.4 that over-generalization due to a lack of string-level information was one of the limitations of UPF-Cobalt that led to an overly permissive behavior. This may be the reason why the performance of the metric is improved with the integration of surface-based metrics.

Finally, Groups III and VI contain the results of the best-performing evaluation metrics from the WMT15 Metrics Task (see Section 4.4.1), as well as the baseline metric BLEU (Papineni et al., 2002) and a strong competitor, Meteor (Denkowski and Lavie, 2014), which we reproduce here for the sake of comparison.

The results demonstrate that fluency features provide useful information regarding the overall translation quality, which is not fully captured by the standard candidate-reference comparison. These features are discriminative when the relationship to the reference does not provide enough information to distinguish between the quality of two alternative candidate translations. For example, it may well be the case that both MT outputs are very different from the human reference, but one constitutes a valid alternative translation, while the other is totally unacceptable.

We further investigate the performance of different groups of fluency features. To that end, in the first place, we separate the LM-based sentence-level features and the extended word-level features described in Section 5.1.2. We train and evaluate the system using these two feature sets and compare the results. As before, the experiments are performed with the WMT15 and WMT14 datasets. In the second place, we isolate the features based on simple word forms and the features based on POS-tag LM and repeat the above procedure.

Our findings are summarized in Table 5.6.

	Average Kendall's τ	
	WMT2014	WMT2015
Fluency-SIMPLE	.143±.013	.142±.011
Fluency-REFINED	.304±.013	.345±.011
Fluency-WORDS	.139±.013	.159±.011
Fluency-POS	.275±.013	.311±.011

Table 5.6: Average Kendall's τ for different groups of fluency features

First, we observe that our fine-grained features (Fluency-REFINED) significantly outperform the sentence-level features (Fluency-SIMPLE) on both WMT14 and WMT15 datasets. In part, this can be explained by the fact that the refined fluency features take into account some alignment information. Specifically, the features that compute backoff behaviour for non-aligned words only have zero values when all the candidate words are aligned. But such cases are fairly rare (they constitute 1.6% and 0.6% of cases on

WMT15 and WMT14 datasets respectively). Our intuition is that the improvement is largely due to including the information on the number of words with low backoff behaviour values that approximates the number of fluency errors in the MT output. We leave for future work a detailed analysis of the impact of individual features on the overall system performance.

Second, as shown in Table 5.6, the features based on POS-tag LM (Fluency-POS) yield a significantly higher Kendall's τ correlation value than word-based features (Fluency-WORDS). This is not surprising, since morphosyntactic representation suffers less from the discrete nature of n-gram language modeling and is more directed to the grammatical ill-formedness of the MT output, which makes it a strong predictor of MT fluency.

5.4 Summary

In this Chapter, we suggested that, in spite of the numerous and varied measurements combined in recent approaches, the fluency of MT output has been overlooked in reference-based evaluation. Existing metrics are largely based on the number of matching words (or POS-tags, or syntactic constructions) between the candidate and reference translations, being thus more measures of sentence similarity than translation quality. In addition, fluency features compensate for an inherent lack of information in reference-based evaluation, when no good match with a reference translation can be found. To further focus on this issue, we designed a set of fine-grained features that takes into account the number of disfluent segments observed in the candidate translation.

In this Chapter we proposed an integral approach to evaluation that takes into account the two defining characteristics of translation quality, its fidelity to the source (still using a reference translation as a proxy) and its compliance to the norms and regularities of target language use.

Chapter 6

ASSESSING REFERENCE BIAS IN MANUAL MT EVALUATION

Throughout this thesis, we have discussed the problem of reference bias in automatic MT evaluation. Automatic evaluation metrics are biased by the reference translation in the sense that the scores generated by the metrics vary considerably if different reference translations are provided (Culy and Riehemann, 2003; Lommel, 2016). This problem has received a lot of attention with considerable work dedicated to the development of evaluation metrics that recognize acceptable differences between MT output and the reference.

The problem of reference bias, however, has barely been considered in the context of manual evaluation. Recall from Section 2.2 that it has currently become common practice to resort to the so called monolingual evaluation that avoids the need to find bilingual speakers. In the monolingual manual evaluation scenario the annotators are target language speakers and they compare the MT to a human translation instead of the source text.

This would be unproblematic if there were one correct solution to the translation task. However, as we have repeatedly seen in this work, the same original text can be translated in many different ways. Therefore, to conduct monolingual evaluation where only one reference translation is provided, one has to assume that human annotators are not influenced by candidate-reference differences if those differences are irrelevant for MT quality. Indeed, in the monolingual reference-based evaluation scenario, human judges are expected to recognize acceptable variations between different translation options and avoid penalizing valid translation choices in MT output even if they happen to be different from the reference translation provided.

In Chapter 3 we showed that the original text can be changed by translators far beyond sheer necessity under the influence of factors that transcend sentence boundaries. We have seen that optional changes in translation include not only grammatical shifts, but also shifts in meaning (Section 3.1.3). Given that the monolingual evaluation task

typically involves assessing the preservation of meaning of the reference translation in the MT output, an interesting question arises: whether manual evaluation is also affected by reference bias. To answer this question, we performed an experiment where the same set of MT outputs is manually assessed using different reference translations and analyzed the discrepancies between the resulting quality scores.

6.1 Experimental Settings

To test if monolingual quality assessment is affected by gold standard human translation, we collected human judgments for the same set of MT outputs evaluated using different reference translations. As control groups, we had different annotators assessing MT outputs using the same reference and using the source segments. Below we present the dataset used in our experiments (Section 6.1.1) and the procedure we followed for collecting human judgments (Section 6.1.2).

6.1.1 Dataset

MT data with multiple references is rare. We used the MTC-P4 Chinese–English dataset, produced by Linguistic Data Consortium (LDC2006T04) (see also Section 4.4.2). The dataset contains 919 source sentences from the news domain, 4 reference translations and MT outputs generated by 10 translation systems. Human translations were produced by four teams of professional translators and included editor's proofreading. All teams used the same translation guidelines, which emphasize faithfulness to the source sentence as one of the main requirements.

We note that even in such a scenario, human translations differ from each other. We measured the average similarity between the four references in the dataset using Meteor (Denkowski and Lavie, 2014). Recall from Section 2.1.1 that Meteor scores range between 0 and 1 and reflect the proportion of similar words occurring in the same order. This metric is normally used to compare the MT output with a human reference, but it can be applied to measure similarity between any two translations. We computed Meteor for all possible combinations between the four available references and took the average score. Even though Meteor covers certain amount of acceptable linguistic variation by allowing for synonym and paraphrase matching, the resulting score is only 0.33, which shows that, not surprisingly, human translations vary substantially.

To make the annotation process feasible given the resources available, we selected a subset of 100 source sentences for the experiment. To ensure variable levels of similarity between the MT and each of the references, we computed sentence-level Meteor scores for the MT outputs using each of the references and selected the sentences with the highest standard deviation between the scores.

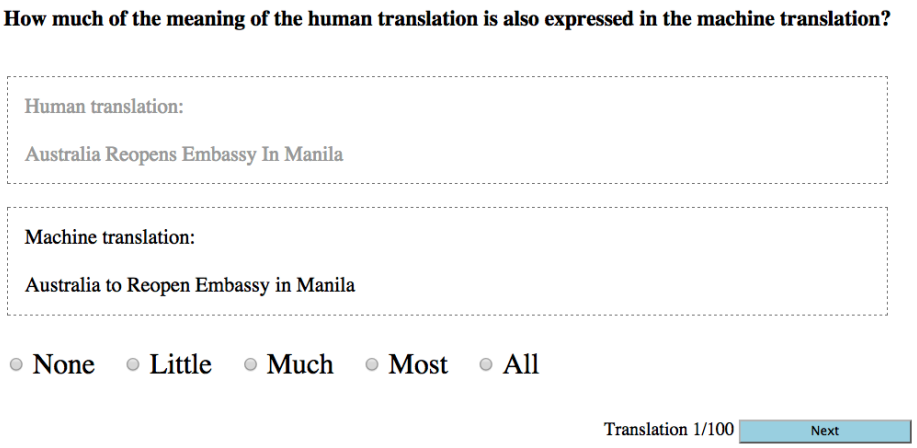


Figure 6.1: Evaluation Interface

6.1.2 Method

We developed a simple online interface to collect human judgments. Our evaluation task was based on the adequacy criterion. Specifically, following a well known formulation of the adequacy evaluation task (Linguistic Data Consortium, 2005; Graham et al., 2015), the judges were asked to estimate how much of the meaning of the human translation was expressed in the MT output (see Figure 6.1). The responses were interpreted on a five-point scale, with the labels in Figure 6.1 corresponding to numbers from 1 ("None") to 5 ("All").

For the main task, judgments were collected from English native speakers who volunteered to participate. They were either professional translators or researchers with a degree in Computational Linguistics, English or Translation Studies. 20 annotators participated in this monolingual task. Each of them evaluated the same set of 100 MT outputs. Our estimates showed that the task could be completed in approximately one hour. The annotators were divided into four groups, corresponding to the four available references. Each group contained five annotators independently evaluating the same set of sentences. Having multiple annotators in each group allowed us to minimize the effect of individual annotators' biases, preferences and expectations.

As a control group, five annotators (native speakers of English, fluent in Chinese or bilingual speakers) performed a bilingual evaluation task for the same MT outputs. In the bilingual task, annotators were presented with an MT output and its corresponding source sentence and asked how much of the meaning of the source sentence was expressed in the MT.

In total, we collected 2,500 human judgments. Both the data and the tool for collecting human judgments are available at `https://github.com/mfomicheva/tradopad.git`.

6.2 Reference Bias

The goal of our experiment was to show that depending on the reference translation used for evaluation, the quality of the same MT output will be perceived differently. We are aware that MT evaluation is a subjective task and certain discrepancies between evaluation scores produced by different annotators are expected simply because of their backgrounds, individual perceptions and expectations regarding translation quality. To show that some differences are related to reference bias and not to the bias introduced by individual annotators, we compared average agreement between annotators who used the same reference vs. annotators who used different references.

Table 6.1 shows the results in terms of standard Cohen (1960) and linearly weighted Cohen (1968) Kappa coefficient (k).[1] We also report one-off version of weighted k, which discards the disagreements unless they are larger than one category. Confidence intervals of the average kappa values for each category of annotators (same reference, different references and source) were computed using bootstrap resampling (Efron and Tibshirani, 1994; Koehn, 2004). This procedure is necessary to estimate how confident we can be that if different annotators were to accomplish the same task, the results would be similar to the ones we obtained. We repeatedly (1000 times) sampled with replacement pairs of annotators of each category, and computed kappa values for each pair and the corresponding average value. We used the 1000 values obtained in this way for each category to compute the final average statistic and the confidence intervals.[2]

As shown in Table 6.1, the agreement between annotators using different references is indeed significantly lower than between annotators using the same reference. Therefore, the same MT outputs systematically receive different scores when different human translations are used for their evaluation.

We suggest that the reason for that may be twofold. On the one hand, annotators may simply choose an easier and more secure evaluation strategy, looking for shared material between the reference and translated sentences. On the one hand, given the optional grammatical or semantical shifts introduced by professional translators with

[1] In MT evaluation, agreement is usually computed using standard k both for ranking different translations and for scoring translations on an interval-level scale. We note, however, that weighted k is more appropriate for scoring, since it allows the use of weights to describe the closeness of the agreement between categories (Artstein and Poesio, 2008) (see also Section 2.2.4).

[2] A slightly different method was reported in (Fomicheva and Specia, 2016) where we used sampling without replacement. As pointed out by Graham et al. (2016), bootstrap resampling is a more standard procedure to be used for the estimation of confidence intervals in our setting. The results reported here and the ones we presented in (Fomicheva and Specia, 2016) are very similar.

Kappa	Diff. ref.	Same ref.	Source
Standard	.161±.002	.188±.004	0.168±.015
Weighted	.328±.002	.361±.004	0.334±.017
One-off	.585±.003	.633±.006	0.594±.026

Table 6.1: Inter-annotator agreement for different-references (Diff. ref.), same-reference (Same ref.) and source-based evaluation (Source)

	Avg
Ref1	1.980
Ref2	2.342
Ref3	2.562
Ref4	2.740

Table 6.2: Average human scores for the groups of annotators using different references

respect to the source text, the annotators may lack contextual and background knowledge to establish the equivalence between the MT output and the reference translation provided.[3]

Agreement between annotators using the source sentences is slightly lower than in the monolingual, same-reference scenario, but it is higher than in the case of the different-reference group. This may be an indication that reference-based evaluation is an easier task for annotators, perhaps because in this case they are not required to shift between languages. Note also that the total number of annotators is considerably smaller in the source category (resulting in larger confidence intervals), as we only had one group of 5 annotators performing the evaluation using source sentences. The comparison between monolingual and bilingual evaluation scenarios may need further investigation.

As a further illustration of reference bias, Table 6.2 shows average evaluation scores for the groups of annotators using different references. Average scores vary considerably across different groups of annotators confirming that MT quality is perceived differently depending on the human translation used as gold-standard.[4]

Finally, to put the results from this Chapter in relation to automatic MT evaluation, Table 6.3 shows Pearson correlation between BLEU scores and human scores produced using each of the available reference translations. Except for Reference 3, the highest correlation is always observed when BLEU scores and human scores were generated using the same reference.

Note also that combinations of reference translations used for human vs. automatic

[3]In line with this discussion, Coughlin (2003) observed in their experiments regarding corpus-level correlation between manual and automatic evaluation that annotators were reluctant to assign MT outputs a maximum score unless they exactly matched the reference. Coughlin (2003) suggested that human annotators may have followed a BLEU-style matching strategy, looking for shared material between the reference and translated sentences and explained that by the tediousness of the task and the annotators' lack of background knowledge in the relevant domains (technical computer domain and Canadian Parliament proceedings in their experiments).

[4]The differences were found to be statistically significant according to Student's t-test (for paired samples).

		BLEU			
		Ref1	Ref2	Ref3	Ref4
Human	Ref1	**0.5331**	0.2651	**0.3957**	0.2984
	Ref2	0.4743	**0.3303**	0.3291	0.3093
	Ref3	0.3321	0.2828	0.3323	0.2633
	Ref4	0.3291	0.2472	0.1743	**0.4200**
	Avg	0.4172	0.2814	0.3079	0.3228

Table 6.3: Pearson correlation between BLEU and human scores produced using different reference translations

evaluation are particularly unfortunate (e.g. the correlation in the case of manual evaluation with Ref3 and automatic evaluation with Ref4 is only 0.1743). In fact, the average correlation varies depending on the reference indicating that different human translations are not equally suitable for automatic evaluation and metric validation purposes.

Evidently, automatic MT evaluation metrics and human annotators are not affected by candidate-reference differences in the same way. In fact, as we have seen throughout this dissertation, penalization of acceptable linguistic variation is one of the reasons the results of automatic evaluation do not correlate well with human judgments at sentence level. However, some of the candidate-reference differences related to linguistic variation in human translation, clearly affect manual evaluation. A very interesting question to explore is what kind of differences between MT and the reference translation introduce reference bias in manual MT evaluation.

Source:	不过这一切都由不得你
Ref*:	However these all totally beyond the control of you.
MT:	But all this is beyond the control of you.
Ref1:	But all this is beyond your control.
Ref2:	However, you cannot choose yourself.
Ref3:	However, not everything is up to you to decide.
Ref4:	But you can't choose that.

Table 6.4: Example of variation between different reference translations (MTC-P4, Chinese–English translation, sentence 396)

Table 6.4 gives an example of linguistic variation in professional human translations and its effect on reference-based MT evaluation. Although all the references carry the same message, linguistic means used by the translators are very different. References 2-4 contain a shift with respect to the meaning of the original sentence. In fact, semantic equivalence between these sentences can only be established on the basis of inference using additional background and contextual information. Annotators are expected to

recognize acceptable variation between the MT and any of the references. However, the average score for this sentence was 3.4 in case of Reference 1, and 2.0, 2.0 and 2.8 in case of the other three references, respectively, which illustrates the bias introduced by the reference translation. It should be noted that the effect of reference bias does not concern only high quality MT. The errors contained in the MT output may be perceived as more or less serious depending on the reference provided.

The results of this experiment are in line with the properties of human translation that we discussed in Chapter 3. Given how different the reference may be from the original text at the level of individual sentences[5], it is not surprising that provided with different translations the annotators assign different scores to the MT output.

6.3 Time Effect

It is well known that the reliability and consistency of human annotation tasks is affected by fatigue (Llorà et al., 2005). As a side question, we examined how this factor may have influenced the evaluation on the impact of reference bias and thus the reliability of our experiment.

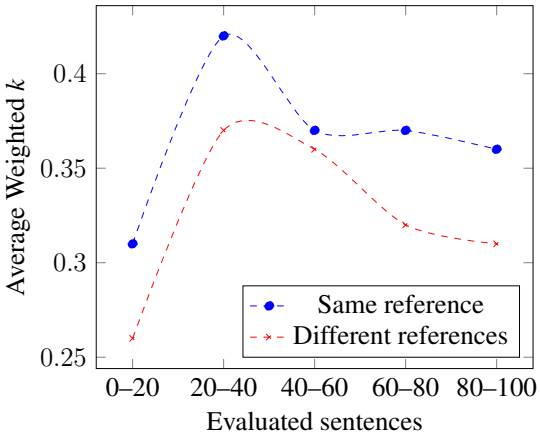


Figure 6.2: Inter-annotator agreement at different stages of evaluation process

We measured inter-annotator agreement for the same-reference and different-reference annotators at different stages of the evaluation process. We divided the dataset in five sets of sentences based on the chronological order in which they were annotated (0-20,

[5]Recall from Chapter 3 that translation equivalence is ultimately established at the level of text, not separate sentences. Therefore, if considered individually original and translated sentences may appear very different. In the example from Table 6.4 without knowing what the text is about, it is very difficult to assess the quality of MT output by comparing it to Ref4. An interesting question for future research is to see whether the results of our experiment would be different if additional context (e.g. the previous and following sentences) were provided.

20-40, ..., 80-100). For each slice of the data we repeated the procedure reported in Section 6.2. Figure6.2 shows the results.

First, we note that the agreement is always higher in the case of same-reference annotators. Second, in the intermediate stages of the task we observe the highest inter-annotator agreement (sentences 20-40) and the smallest difference between the same-reference and different-reference annotators (sentences 40-60). This seems to indicate that the effect of reference bias is minimal half-way through the evaluation process. In other words, when the annotators are already acquainted with the task but not tired yet, they are able to better recognize meaning-preserving variation between different translation options.

To further investigate how fatigue affects the evaluation process, we tested the variability of human scores in different (chronological) slices of the data. We again divided the data in five sets of sentences and calculated standard deviation between the scores in each set. We repeated this procedure for each annotator and averaged the results. As can be seen in Figure 6.3, the variation between the scores is lower in the last stages of the evaluation process. This could mean that towards the end of the task the annotators tend to indiscriminately give similar scores to any translation, making the evaluation less informative.

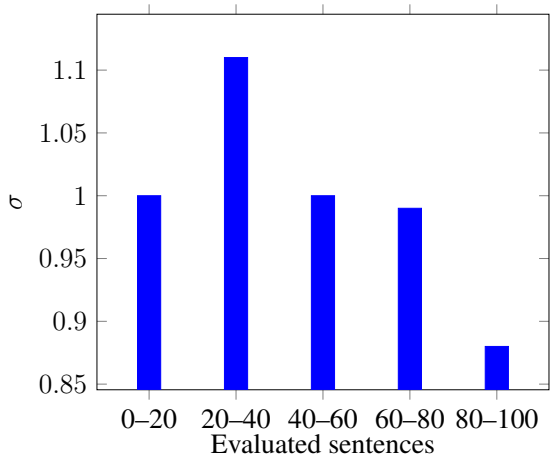


Figure 6.3: Average standard deviations between human scores for all annotators at different stages of evaluation process

6.4 Summary

We examined the effect of reference bias on monolingual MT evaluation. We compared the agreement between the annotators who used the same human reference translation and those who used different reference translations. We demonstrated that in addition to the inevitable bias introduced by different annotators, monolingual evaluation is systematically affected by the reference provided. Annotators consistently assign different scores to the same MT outputs when a different human translation is used as gold-standard.

Quality assessment is instrumental in the development and deployment of MT systems. If evaluation is to be objective and informative, its purpose must be clearly defined. The same sentence can be translated in many different ways. Using a human reference as a proxy for the source sentence, we evaluate the similarity of the MT to a particular reference, which does not necessarily reflect how well the contents of the original is expressed in the MT or how suitable it is for a given purpose.

An alternative to reference-based evaluation is not easy to find. One possibility is to employ the original text instead of the reference, but this may be infeasible in practice given the difficulty of finding bilingual speakers. Other possibilities include controlling the type of reference used for evaluation, for instance, by working with certain text types where the translation priority is the preservation of the original as closely as possible. Alternatively, translation guidelines can be established for the creation of reference translations specifying that the translated text must be as close as possible to the original, with the correspondences holding at the lowest level possible. In conclusion, the requirements to the gold standard for MT evaluation is an open question that needs to be answered taking into consideration the evaluation settings (purpose of evaluation, type of MT systems, domain, etc.), as well as the priorities in MT development (e.g. the relative importance of fluency and adequacy aspects).

Chapter 7

CONCLUSIONS

This final chapter provides a summary of the contributions of our work, as well as the directions for future research. Section 7.1 presents the conclusions. In Section 7.2 the contributions are summarized. Finally, Section 7.3 outlines directions for future work.

7.1 General Conclusions

In this dissertation we discussed the problems of MT evaluation originating from the use of human translation as a benchmark against which MT is compared. Our main goals were, first, to assess the effects of using human translation on MT evaluation, and second, to improve the accuracy of automatic reference-based MT evaluation.

In regards to the first goal, this work has provided a detailed discussion of the properties of human translation that questions the appropriateness of using it as reference for the purposes of MT evaluation. We established that the presence of translation shifts in the reference introduces noise in automatic MT evaluation, as the differences related to translation errors and to optional changes in human translation are penalized in the same way. The number of shifts in the reference translation depends on the text type and genre, and the underlying translation strategy. Using the reference as benchmark makes the results of evaluation less stable across domains and use cases, since metric scores depend not only on the actual quality of the MT, but also on how difficult it is to match the reference in this particular setting.

Our experiments on automatic generation of additional references reported in Section 3.2 showed that evaluation accuracy can be improved using close translation variants. The analysis of the results also brought into our view the fact that some of the optional changes present in human translation can be found in MT. Given the potential of statistical MT to infer knowledge from data and generate more "human-like" output, using close translation alone for reference-based evaluation may be too limiting. In this sense, a set of references varying in terms of the distance with respect to the original is

an ideal setting that could provide information regarding the levels of quality that can be attained by MT systems based on different strategies (e.g. rule-based vs. statistical MT).

This, however, is difficult to achieve in practice. Instead, to address our second objective, we proposed a shift in perspective from focusing on reference similarity to assessing MT quality while still using the available reference as one of the sources of information. Specifically, rather than aspects of similarity, we suggested to concentrate on different types of candidate-reference differences and search for criteria that would allow to distinguish between acceptable variations and deviations induced by various types of MT errors. The approach we developed throughout Chapters 4 and 5 successfully addressed the improvement of current evaluation metrics. Our approach is crucially based on how to weight candidate-reference differences to reflect their impact on perceived MT quality. The results from Chapter 5 confirmed the benefits of introducing various techniques to this end: context evidence for candidate-reference alignment, word embeddings based representation to increase lexical coverage, validation of lexical matches through syntactic context and weighting of context differences based on the functions of the words involved, as well as the number of their syntactic dependents.

The analysis of the results highlighted various benefits of the approach proposed. On the one hand, introducing additional constraints by requiring that matching words play equal or similar roles in the corresponding sentences allowed to lower the restrictions for lexical similarity and account for contextual synonyms. On the other hand, the results showed that the impact of candidate-reference differences on human perception of translation quality is better approximated taking into account not only the type of underlying translation error – word order, missing word, added word, etc. – but also the relevance of the words involved for the interpretation of the sentence, measured in our experiments in terms of the number of syntactic dependents and the type of syntactic function. Finally, in line with the idea of moving the focus from linguistic levels of similarity to different aspects of quality, the results reported in Chapter 5 clearly demonstrated the benefits of including fluency-oriented features into reference-based evaluation.

The last part of this work presented in Chapter 6 was dedicated to manual assessment of MT quality. We demonstrated that in addition to the inevitable bias introduced by individual preferences and expectations of the annotators, monolingual evaluation is systematically affected by the reference translation. Annotators consistently assign different scores to the same MT outputs when a different human translation is used as gold-standard. The requirements to the gold standard for MT evaluation is an open question that needs to be answered taking into consideration the evaluation settings (purpose of evaluation, type of MT systems, domain, etc.), as well as the priorities in MT development (e.g. the relative importance of fluency and adequacy aspects). However, it should be stressed that the consistency and reliability of manual evaluation is criti-

cal for a meaningful comparison between different automatic evaluation strategies and its analysis and improvement must be a part of the research in automatic MT evaluation. In fact, in our view existing automatic evaluation metrics are close to the limits of their performance set by the nature of reference translation and the reliability of manual evaluation.

7.2 Contributions

The main contributions of this dissertation concern a theoretical discussion and an empirical investigation of the impact of the differences between MT and human translation on the results of automatic and manual evaluation of MT quality. The results of this thesis have been published in 5 peer-reviewed publications (Fomicheva et al., 2015a,b; Fomicheva and Bel, 2016; Fomicheva et al., 2016; Fomicheva and Specia, 2016), in which we describe the approaches to automatic evaluation developed during this work, as well as our investigation of the reference bias in manual evaluation. The following are the contributions that resulted from the work presented in this thesis:

- *Analysis of the challenges of reference-based evaluation from the perspective of translation studies.* We discussed the concepts of translation equivalence, translation universals and translation shifts. We suggested that the priority that is typically given by translators to pragmatic equivalence, as well as the impact of the factors inherent to the human translation process results in optional translation shifts, i.e. deviations from a close translation variant.
- *Modeling of prototypical syntactic shifts in English–Spanish translation.* We designed a prototype paraphrase generation system based on a set of syntactic transformation rules that model structural changes contained in the reference translation. The system was expected to paraphrase the initial reference "undoing" optional translation shifts that it could contain.
- *Analysis of the impact of syntactic translation shifts on the results of automatic reference-based MT evaluation.* We employed automatic paraphrase generation to create close translation alternatives to be used as additional references. We compared the results of automatic evaluation with one human reference and with multiple generated references confirming that MT outputs obtain low scores when compared to a shifted reference and higher scores when compared to automatically generated paraphrases in which the potential shifts were eliminated. We also compared the correlation with human judgments for these two evaluation scenarios and found a significant improvement in multi-reference scenario, confirming that MT outputs assigned lower scores by the metrics in cases of optional shifts obtained higher scores from human judgments based on the comparison of the MT output with the source sentence.

- *Use of advanced monolingual alignment techniques with context evidence for the purposes of MT evaluation.* We demonstrated that the use of context evidence for candidate-reference alignment alleviates the problem of spurious matches in automatic MT evaluation.
- *Increase of the coverage of acceptable linguistic variation in reference-based MT evaluation.* On the one hand, we increased lexical coverage for candidate-reference alignment through the use of distributional similarity measures over vector word representations. We integrated an additional lexical similarity component based on word embeddings into the MWA aligner. We showed that using distributional similarity for candidate-reference alignment, in combination with a penalty for the differences in word context results in an increase in the correlation with human judgments. On the other hand, we integrated into evaluation a mapping between semantically equivalent dependency functions increasing the coverage of acceptable linguistic variation at syntactic level.
- *Use of syntactic context to estimate the impact of candidate-reference word-level differences on sentence-level MT quality as perceived by human judges.* We modeled the effect of candidate-reference differences on perceived translation quality taking various measurements on the pairs of aligned words: (a) number of syntactically related matching words, (b) overall number of syntactic dependents, and (c) type of syntactic function (arguments, modifiers, specifier).
- *Development of a new automatic evaluation metric, UPF-Cobalt, that contains the features described above and attains a highly competitive performance on various evaluation datasets.* The metric and the code is publicly available for research purposes at `https://github.com/mfomicheva/upf-cobalt`.
- *New method for a detailed meta-evaluation of evaluation metrics consisting in the analysis of metric performance with MT outputs of different levels of underlying quality.* The analysis revealed some specific limitations of our evaluation metric, as well as a general observation of a decrease in metric performance on low quality MT output.
- *Integration of fluency-oriented features into reference-based MT evaluation.* We established that one of the fundamental problems in reference-based evaluation is a complete lack of information regarding the characteristics of MT words that are not matched with the reference. As a strategy to compensate for this lack of information we proposed to use fluency-oriented features. In addition to LM-based features widely used for reference-free evaluation, we designed a more detailed representation of MT fluency that takes into account the number of disfluent segments in the MT output. To include the fluency aspect in reference-based MT evaluation, we designed a feature-based approach to evaluation that combined the features representing translation fluency with adequacy-oriented features derived from UPF-Cobalt. We further boosted performance by combining the scores from

reference-based metrics (BLEU, Meteor and UPF-Cobalt) with fluency-oriented features. On pair with the DPMFComb metric (Yu et al., 2015), our approach attained the best results at the WMT16 Metrics Task (Bojar et al., 2016).

- *Demonstration of the impact of reference bias on manual MT evaluation.* We showed that human evaluation is affected by reference bias, i.e. human annotators systematically assign different scores to the same MT output if a different reference translation is provided for evaluation.
- *Construction of monolingual evaluation dataset with human judgments elicited with the use of different reference translations.* The dataset is publicly available at `https://github.com/mfomicheva/tradopad`.

7.3 Future Work

The results of this dissertation set various new lines of research. Concerning our experiment on close translation generation, we plan to repeat the experiments using MT systems based on different strategies (e.g. rule-based MT vs. statistical MT). We expect that the benefits of using close translations for evaluation would be more pronounced in the case of rule-based MT. Along the same lines, an interesting direction for further research concerns automatic selection or filtering of sentences with human reference translations that are not suitable for automatic evaluation. This could be achieved, for example, based on a series of features reflecting source-reference similarity.

Another line of further work concerns the analysis of the performance of our automatic evaluation metrics. We plan to test them for MT optimization and compare the results with BLEU and Meteor, which are commonly used for this purpose. We expect, in general, that a varied set of metrics used for optimization will bring about better quality of MT output. Furthermore, we would like to test the performance of our evaluation strategies on languages other than English, assessing to what extent the performance of the metrics is affected by the decrease in the reliability of syntactic parsing. Finally, we would like to extend on our analysis of metric performance on different levels of underlying MT quality to other evaluation metrics and datasets, as it provides detailed information regarding the limitations of the metrics and can suggest further ways of improvement.

Regarding further development of our evaluation metrics, first, we would like to design a light version of UPF-Cobalt using linear word context instead of dependency-based representation. Second, an interesting idea for addressing acceptable variation with distributed word representations is to explore context substitution methods that identify meaning-preserving substitutes for a target word instance in a given sentential context (Melamud et al., 2015). An evaluation score can then be computed based on the extent to which the MT words are considered suitable substitutes for reference words. Finally, feature selection methods can be used to improve our feature-based approach

presented in Chapter 5.

We would also like to use the alignment from UPF-Cobalt as well as the information on lexical similarity and context penalty for automatic error annotation and compare the results with human annotation on available datasets (e.g. (Lommel et al., 2014; Blain et al., 2016)). The results of automatic error analysis can be used to generate metric scores based on the types of errors. The relation of the types of MT errors and the perceived MT quality constitutes an interesting line of further research.

Finally, regarding manual evaluation, a very interesting direction for future research is to analyze what kind of acceptable differences between MT and the reference translation introduce reference bias in manual MT evaluation, as it is evident that automatic MT evaluation metrics and human annotators are not affected by the differences in the same way.

Bibliography

Lars Ahrenberg. Codified Close Translation as a Standard for MT. In *Proceedings of 10th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 13–22, 2005.

Lars Ahrenberg and Magnus Merkel. Correspondence Measures for MT evaluation. In *Proceedings of the Second International Conference on Linguistic Resources and Evaluation (LREC)*, pages 1255–1261, 2000.

Ahmet Aker, Frederic Blain, Andres Duque, Marina Fomicheva, Jurica Seva, Kashif Shah, and Daniel Beck. USFD at SemEval-2016 Task 1: Putting different State-of-the-Arts into a Box. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2016)*, pages 609–613, 2016.

Joshua Albrecht and Rebecca Hwa. Regression for Sentence-Level MT Evaluation with Pseudo-References. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 45, pages 296–303, 2007.

Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Felisa Verdejo. The Contribution of Linguistic Features to Automatic Machine Translation Evaluation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 306–314, 2009.

Ron Artstein and Massimo Poesio. Inter-coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. Goodness: A Method for Measuring Machine Translation Confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL): Human Language Technologies*, pages 211–219, 2011.

Mona Baker. Corpus Linguistics and Translation Studies: Implications and Applications. In *Text and technology: In honour of John Sinclair*, pages 233–250. John Benjamins Publishing Company, 1993.

Mona Baker. Corpus-based Translation Studies: The Challenges that Lie Ahead. *Benjamins Translation Library*, 18:175–186, 1996.

Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Association for Computational Linguistics (ACL) Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, volume 29, pages 65–72, 2005.

Colin Bannard and Chris Callison-Burch. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 597–604, 2005.

Marco Baroni and Alessandro Lenci. Distributional Memory: A General Framework for Corpus-based Semantics. *Computational Linguistics*, 36(4):673–721, 2010.

Marco Baroni, Dinu Georgiana, and Germán Kruszewski. Don't Count, Predict! A Systematic Comparison of Context-counting vs. Context-predicting Semantic Vectors. In *ACL (1)*, pages 238–247, 2014.

Alberto Barrón-Cedeño, Marta Vila, M Antònia Martí, and Paolo Rosso. Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *Computational Linguistics*, 39(4):917–947, 2013.

Viktor Becher. *Explicitation and implicitation in translation: A corpus-based study of English-German and German-English translations of business texts*. PhD thesis, Hamburg, Universität Hamburg, 2011.

Rahul Bhagat and Eduard Hovy. What is a paraphrase? *Computational Linguistics*, 39 (3):463–472, 2013.

Steven Bird. NLTK: the Natural Language Toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72, 2006.

Frédéric Blain, Jean Senellart, Holger Schwenk, Mirko Plitt, and Johann Roturier. Qualitative Analysis of Post-Editing for High Quality Machine Translation. *MT Summit XIII: the Thirteenth Machine Translation Summit [organized by the] Asia-Pacific Association for Machine Translation (AAMT)*, pages 164–171, 2011.

Frédéric Blain, Varvara Logacheva, and Lucia Specia. Phrase level segmentation and labelling of machine translation errors. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2240–2245, 2016.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence Estimation for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, 2004.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar F Zaidan. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11. Association for Computational Linguistics, 2011.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, 2014.

Ondrej Bojar, Yvette Graham, and AKM Stanojevic. Results of the WMT16 Metrics Shared Task. In *Proceedings of the First Conference on Machine Translation*, pages 199–231, 2016.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September 2015.

Chris Brockett. Aligning the RTE 2006 Corpus. Technical report, Microsoft Research, 2007.

Sabine Buchholz and Erwin Marsi. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164, 2006.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to Rank Using Gradient Descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96, 2005.

Chris Callison-Burch. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 196–205, 2008.

Chris Callison-Burch and Miles Osborne. Re-evaluating the Role of BLEU in Machine Translation Research. In *In Proceedings of the European Association for Computational Linguistics (EACL)*, pages 249–256, 2006.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, 2007.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F Zaidan. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, 2010.

John Catford. *A Linguistic Theory of Translation: An Essay in Applied Linguistics*. Oxford: Oxford University Press, 1965.

Eugene Charniak, Kevin Knight, and Kenji Yamada. Syntax-based language models for statistical machine translation. In *Proceedings of MT Summit IX*, pages 40–46, 2003.

Boxing Chen and Colin Cherry. A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, 2014.

Stanley F Chen and Joshua Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics, 1996.

David Chiang, Yuval Marton, and Philip Resnik. Online large-margin training of syntactic and structural translation features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 224–233, 2008.

Kenneth W Church and Eduard H Hovy. Good Applications for Crummy Machine Translation. *Machine Translation*, 8(4):239–258, 1993.

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Jacob Cohen. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological bulletin*, 70(4):213–220, 1968.

Elisabet Comelles, Jordi Atserias, Victoria Arranz, and Irene Castellón. VERTa: Linguistic Features in MT Evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 3944–3950, 2012.

Deborah Coughlin. Correlating Automated and Human Assessments of Machine Translation Quality. In *Proceedings of the MT Summit IX*, pages 63–70, 2003.

Christopher Culy and Susanne Z Riehemann. The Limits of N-gram Translation Evaluation Metrics. In *Proceedings of the MT Summit IX*, pages 71–78, 2003.

Lea Cyrus. Building a Resource for Studying Translation Shifts. In *Proceedings of the Fifth International Conference on Linguistic Resources and Evaluation (LREC)*, pages 1240–1245, 2006.

Lea Cyrus. Old Concepts, New Ideas: Approaches to Translation Shifts. *MonTI. Monografías de Traducción e Interpretación*, (1):87–106, 2009.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190. 2006.

Marie-Catherine De Marneffe and Christopher D Manning. Stanford Typed Dependencies Manual. Technical report, Stanford University, 2008.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454. Genoa, 2006.

Michael Denkowski and Alon Lavie. Meteor-next and the meteor paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342, 2010a.

Michael Denkowski and Alon Lavie. Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks. In *Proceedings of the Ninth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, 2010b.

Michael Denkowski and Alon Lavie. Exploring normalization techniques for human judgments of machine translation adequacy collected using amazon mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 57–61, 2010c.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, 2011.

Michael Denkowski and Alon Lavie. Meteor Universal: Language Specific Translation Evaluation for any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, 2014.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, 2002.

Monika Doherty. *Structural propensities: translating nominal word groups from English into German*, volume 65. John Benjamins Publishing, 2006.

Bonnie Dorr, Joseph Olive, John McCary, and Caitlin Christianson. Machine Translation Evaluation and Optimization. In *Handbook of Natural Language Processing and Machine Translation*, pages 745–843. Springer, 2011.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. CRC Press, 1994.

Mireia Farrús, Marta Ruiz Costa-Jussà, José Bernardo Mariño Acebal, and José Adrián Rodríguez Fonollosa. Linguistic-based Evaluation Criteria to Identify Statistical Machine Translation Errors. In *14th Annual Conference of the European Association for Machine Translation*, pages 167–173, 2010.

Marcello Federico, Matteo Negri, Luisa Bentivogli, Marco Turchi, and FBK-Fondazione Bruno Kessler. Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1643–1653, 2014.

Mariano Felice and Lucia Specia. Linguistic Features for Quality Estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 96–103, 2012.

C. Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 363–370, 2005.

Mark Fishel, Rico Sennrich, Maja Popović, and Ondřej Bojar. Terrorcat: a Translation Error Categorization-Based MT Quality Metric. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 64–70, 2012.

Marina Fomicheva and Núria Bel. Using Contextual Information for Machine Translation Evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 2755–2761, 2016.

Marina Fomicheva and Lucia Specia. Reference Bias in Monolingual Machine Translation Evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 77–82, 2016.

Marina Fomicheva, Núria Bel, and Iria da Cunha. Neutralizing the Effect of Translation Shifts on Automatic Machine Translation Evaluation. In *Computational Linguistics and Intelligent Text Processing*, pages 596–607. Springer, 2015a.

Marina Fomicheva, Núria Bel, Iria da Cunha, and Anton Malinovskiy. UPF-Cobalt Submission to WMT15 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 373–379, 2015b.

Marina Fomicheva, Núria Bel, Lucia Specia, Iria da Cunha, and Anton Malinovskiy. CobaltF: A Fluent Metric for MT Evaluation. In *Proceedings of the First Conference on Statistical Machine Translation (WMT)*, volume 2, pages 483–490, 2016.

Lauren Friedman, Haejoong Lee, and Stephanie Strassel. A Quality Control Framework for Gold Standard Reference Translations: The Process and Toolkit Developed for GALE. In *Proceedings of the Thirtieth International Conference on Translating and the Computer*, 2008.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 758–764, 2013.

Jesús Giménez. *Empirical Machine Translation and its Evaluation*. PhD thesis, Universitat Politècnica de Barcelona, 2008.

Jesús Giménez and Lluís Màrquez. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, 1 (94):77–86, 2010.

Jesús Giménez and Lluís Màrquez. Linguistic Measures for Automatic Machine Translation Evaluation. *Machine Translation*, 24(3-4):209–240, 2010.

Yvette Graham and Timothy Baldwin. Testing for Significance of Increased Correlation with Human Judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, 2014.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 33–41, 2013.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. Accurate Evaluation of Segment-Level Machine Translation Metrics. In *Proceedings of NAACL-HLT*, pages 1183–1191, 2015.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. Is all that Glitters in MT Quality Estimation really Gold Standard? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, 2016.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1066–1072, 2015.

Ernst-August Gutt. *Translation and Relevance: Cognition and Context*. Routledge, 2014.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Using Discourse Structure Improves Machine Translation Evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 687–698, 2014.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Machine Translation Evaluation with Neural Networks. *Computer Speech & Language*, 2016.

Nizar Habash and Ahmed Elkholy. SEPIA: surface span extension to syntactic dependency precision-based MT evaluation. In *Proceedings of the NIST metrics for machine translation workshop at the association for machine translation in the Americas conference, AMTA-2008. Waikiki, HI*, 2008.

Hans Halteren van. Source Language Markers in EUROPARL Translations. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 937–944. Association for Computational Linguistics, 2008.

Christian Hardmeier. *Discourse in Statistical Machine Translation*. PhD thesis, Uppsala Universitet, 2014.

Zellig S. Harris. Distributional Structure. *Word*, 10(2-3):146–162, 1954.

Yifan He, Jinhua Du, Andy Way, and Josef van Genabith. The DCU dependency-based metric in WMT-MetricsMATR 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 349–353, 2010.

James Holmes. The Name and Nature of Translation Studies. *Translated! Papers on literary translation and translation studies*, pages 67–80, 1988.

Mark Hopkins and Jonathan May. Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1352–1362, 2011.

Juliane House. Translation Quality Assessment: Linguistic Description versus Social Evaluation. *Meta: Journal des traducteurs*, 46(2):243–257, 2001.

Juliane House. *Routledge Encyclopedia of Translation Studies*, chapter Quality, pages 222–225. 2009.

Eduard Hovy, Margaret King, and Andrei Popescu-Belis. Principles of context-based machine translation evaluation. *Machine Translation*, 17(1):43–75, 2002.

William John Hutchins and Harold Somers. *An Introduction to Machine Translation*. Academic Press London, 1992.

Zahurul Islam and Alexander Mehler. Customization of the Europarl Corpus for Translation Studies. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation LREC-2012*, pages 2505–2510, 2012.

Jeremy G Kahn, Matthew Snover, and Mari Ostendorf. Expected dependency pair match: predicting translation quality with expected syntactic structure. *Machine Translation*, 23(2-3):169–179, 2009.

David Kauchak and Regina Barzilay. Paraphrasing for Automatic Evaluation. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 455–462. Association for Computational Linguistics, 2006.

Margaret King, Andrei Popescu-Belis, and Eduard Hovy. FEMTI: creating and using a framework for MT evaluation. In *Proceedings of MT Summit IX, New Orleans, LA*, pages 224–231, 2003.

Katrin Kirchhoff, Daniel Capurro, and Anne M Turner. A Conjoint Analysis Framework for Evaluating User Preferences in Machine Translation. *Machine Translation*, 28(1): 1–17, 2014.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395, 2004.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, pages 79–86, 2005.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009.

Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121. Association for Computational Linguistics, 2006.

Werner Koller. Equivalence in Translation Theory. In Andrew Chesterman, editor, *Readings in Translation Theory*, pages 99–104. Helsinki: Oy Finn Lectura Ab., 1989.

David Kurokawa, Cyril Goutte, Pierre Isabelle, et al. Automatic detection of translated text and its impact on machine translation. *Proceedings of MT-Summit XII*, pages 81–88, 2009.

J Richard Landis and Gary G Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, pages 159–174, 1977.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Adapting Translation Models to Translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265. Association for Computational Linguistics, 2012.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. CDER: Efficient MT Evaluation Using Block Movements. In *Proceedings of the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

Omer Levy and Yoav Goldberg. Dependency-Based Word Embeddings. In *Proceedings of ACL (2)*, pages 302–308, 2014.

Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. An end-to-end discriminative approach to machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 761–768. Association for Computational Linguistics, 2006.

Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605. Association for Computational Linguistics, 2004.

Linguistic Data Consortium. Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations. Technical report, Linguistic Data Consortium, 2005.

Ding Liu and Daniel Gildea. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32, 2005.

Ding Liu and Daniel Gildea. Source-language features and maximum correlation training for machine translation evaluation. In *Proceedings of HLT-NAACL*, volume 7, pages 41–48, 2007.

Xavier Llorà, Kumara Sastry, David E Goldberg, Abhimanyu Gupta, and Lalitha Lakshmi. Combating User Fatigue in iGAs: Partial Ordering, Support Vector Machines, and Synthetic Fitness. In *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation*, pages 1363–1370, 2005.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. Fully automatic semantic mt evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 243–252. Association for Computational Linguistics, 2012.

Arle Lommel. Blues for BLEU: Reconsidering the Validity of Reference-Based MT Evaluation. In *Proceedings of the LREC 2016 Workshop Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 63–70, 2016.

Arle Lommel, Aljoscha Burchardt, Maja Popovic, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. Using a new analytic measure for the annotation and analysis of mt errors on real data. In *Proceedings of EAMT*, 2014.

Ngoc-Quang Luong, Laurent Besacier, and Benjamin Lecouteux. Towards Accurate Predictors of Word Quality for Machine Translation: Lessons Learned on French–English and English–Spanish Systems. *Data & Knowledge Engineering*, 96:32–42, 2015.

Bill MacCartney, Trond Grenager, Marie-Catherine Marneffe, Daniel Cer, and Christopher D Manning. Learning to recognize features of valid textual entailments. In *Proceedings of the main conference of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 41–48, 2006.

Matouš Macháček and Ondřej Bojar. Approximating a deep-syntactic metric for mt evaluation and tuning. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 92–98, 2011.

Matouš Macháček and Ondřej Bojar. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–10, 2014a.

Matouš Macháček and Ondřej Bojar. Results of the WMT14 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–10, 2014b.

Nitin Madnani and Bonnie J Dorr. Generating Targeted Paraphrases for Improved Translation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):40, 2013.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie J Dorr. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 120–127. Association for Computational Linguistics, 2007.

Benjamin Marie and Marianna Apidianaki. Alignment-based Sense Selection in METEOR and the RATATOUILLE Recipe. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 385–391, 2015.

Montserrat Marimon. The Spanish DELPH-IN Grammar. *Language Resources and Evaluation*, 47(2):371–397, 2013.

Montserrat Marimon, Nuria Bel, and Lluís Padró. Automatic selection of hpsg-parsed sentences for treebank construction. *Computational Linguistics*, 40(3):523–531, 2014.

Dennis N Mehay and Chris Brew. BLEUÂTRE: Flattening Syntactic Dependencies for MT Evaluation. *TMI 2007*, page 122, 2006.

Dan Melamed, Ryan Green, and Joseph P Turian. Precision and Recall of Machine Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 61–63. Association for Computational Linguistics, 2003.

Oren Melamud, Omer Levy, Ido Dagan, and Israel Ramat-Gan. A Simple Word Embedding Model for Lexical Substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 1–7, 2015.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

Melanie Mitchell. *An Introduction to Genetic Algorithms*. MIT press, 1998.

Roberto Navigli and Simone Paolo Ponzetto. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics, 2010.

Eugene Nida. Principles of Correspondence. *The Translation Studies Reader*, 3:141–155, 1964/2000.

Sonja Nießen, Franz Josef Och, Gregor Leusch, Hermann Ney, et al. An evaluation tool for machine translation: Fast evaluation for mt research. In *Proceedings of the LREC Conference*, 2000.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135, 2007.

Sharon O'Brien. *Cognitive explorations of translation*. Bloomsbury Publishing, 2011.

FJ Och, D Gildea, S Khudanpur, A Sarkar, K Yamada, A Fraser, S Kumar, L Shen, D Smith, K Eng, V Jain, Z Jin, and D Radev. Final Report of Johns Hopkins 2003 Summer Workshop on Syntax for Statistical Machine Translation. Technical Report. Technical report, Johns Hopkins University, 2003.

Joseph Olive. Global autonomous language exploitation (GALE). *DARPA/IPTO Proposer Information Pamphlet*, 2005.

Karolina Owczarzak, Declan Groves, Josef Van Genabith, and Andy Way. Contextual bitext-derived paraphrases in automatic mt evaluation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 86–93, 2006.

Karolina Owczarzak, Josef Van Genabith, and Andy Way. Dependency-based automatic evaluation for machine translation. In *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 80–87, 2007.

Sebastian Padó, Daniel Cer, Michel Galley, Dan Jurafsky, and Christopher D Manning. Measuring machine translation quality as semantic equivalence: A metric based on entailment features. *Machine Translation*, 23(2-3):181–193, 2009.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

Karl Pearson. *The Life, Letters and Labours of Francis Galton*, volume 2. CUP Archive, 1924.

Maja Popovic. CHRF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, 2015.

Maja Popović and Hermann Ney. Syntax-oriented evaluation measures for machine translation output. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 29–32. Association for Computational Linguistics, 2009.

Maja Popović and Hermann Ney. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688, 2011.

Maja Popovic, Arle Lommel, Aljoscha Burchardt, Eleftherios Avramidis, and Hans Uszkoreit. Relations between different types of post-editing operations, cognitive effort and temporal effort. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT14)*, pages 191–198, 2014.

Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

Ying Qin and Lucia Specia. Insight into Multiple References in an MT Evaluation Metric. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 131–140. Springer, 2015.

Christopher Quirk. Training a sentence-level machine translation confidence measure. In *Proceedings of LREC*, 2004.

Sylvain Raybaud, David Langlois, and Kamel Smaïli. "this sentence is wrong." detecting errors in machine-translated sentences. *Machine Translation*, 25(1):1–34, 2011.

Florence Reeder, K Miller, K Doyon, and J White. The naming of things and the confusion of tongues. In *Proceedings of the 4th ISLE Evaluation Workshop, MT Summit VIII. Santiago de Compostela, Spain*, 2001.

Katharina Reiss. *Translation Criticism – Potentials and Limitations: Categories and Criteria for Translation Quality Assessment*. Routledge, 2014.

Helmut Schmid. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer, 1999.

Holger Schwenk. Continuous Space Language Models. *Computer Speech & Language*, 21(3):492–518, 2007.

Gilles Sérasset. DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web*, 6(4):355–361, 2015.

Christophe Servan, Alexandre Bérard, Zied Elloumi, Hervé Blanchon, and Laurent Besacier. Word2Vec vs DBnary: Augmenting METEOR using Vector Representations or Lexical Resources? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1159–1168, 2016.

Kashif Shah, Trevor Cohn, and Lucia Specia. An Investigation on the Effectiveness of Features for Translation Quality Estimation. In *Proceedings of the Machine Translation Summit*, volume 14, pages 167–174, 2013.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. Hypernyms under Siege: Linguistically-motivated Artillery for Hypernymy Detection. In *Proceedings of EACL 2017*, pages 1–8, 2017.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, 2006.

Matthew Snover, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. Fluency, Adequacy, or HTER?: Exploring Different Human Judgments with a Tunable MT Metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268. Association for Computational Linguistics, 2009a.

Matthew G Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3):117–127, 2009b.

Xingyi Song, Trevor Cohn, and Lucia Specia. BLEU deconstructed: Designing a better MT evaluation metric. *International Journal of Computational Linguistics and Applications*, 4(2):29–44, 2013.

Charles Spearman. The Proof and Measurement of Associati on Between Two Rings. *American Journal of Psychology*, (15):72–101, 1904.

Lucia Specia. Exploiting Objective Annotations for Measuring Translation Post-Editing Effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 73–80, 2011.

Lucia Specia and Jesús Giménez. Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation. In *The Ninth Conference of the Association for Machine Translation in the Americas*, 2010.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. Estimating the Sentence-level Quality of Machine Translation Systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37, 2009.

Lucia Specia, Nicola Cancedda, and Marc Dymetman. A dataset for assessing machine translation evaluation metrics. In *Proceedings of LREC*, pages 3375–3378, 2010a.

Lucia Specia, Dhwaj Raj, and Marco Turchi. Machine Translation Evaluation versus Quality Estimation. *Machine Translation*, 24(1):39–50, 2010b.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. Multi-level Translation Quality Prediction with QuEst++. In *53rd Annual Meeting of the Association for Computational Linguistics and Seventh International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: System Demonstrations*, pages 115–120, 2015.

Miloš Stanojevic and Khalil Sima'an. BEER: BEtter evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, 2014.

Miloš Stanojevic and Khalil Sima'an. BEER 1.1: ILLC UvA Submission to Metrics and Tuning Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 396–401, 2015.

Miloš Stanojevic, Amir Kamran, Philipp Koehn, and Ondrej Bojar. Results of the WMT15 Metrics Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, 2015.

Erich Steiner. Grammatical metaphor in translation–some methods for corpus-based investigations1. *Language and Computers*, 39(1):213–228, 2002.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence. *Transactions of the ACL*, 2:219–230, 2014.

Izabela Szymańska. *Mosaics: A construction-grammar-based approach to translation*. Wydawnictwo Naukowe" Semper", 2011.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved Semantic Representations from Tree-Structured Long Short-Term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and the 7th International Joint Conference on Natural Language Processing*, pages 1556–1566, 2015.

Irina P Temnikova. Cognitive evaluation approach for a controlled language post–editing experiment. In *Proceedings of LREC*, pages 3485–3490, 2010.

Kapil Thadani, Scott Martin, and Michael White. A Joint Phrasal and Dependency Model for Paraphrase Alignment. In *Proceedings of COLING 2012: Posters*, 2012.

Antonio Toral, Sudip Naskar, Federico Gaspari, and Declan Groves. DELiC4MT: A Tool for Diagnostic MT Evaluation over User-defined Linguistic Phenomena. *The Prague Bulletin of Mathematical Linguistics*, 98:121–131, 2012.

Gideon Toury. Probabilistic Explanations in Translation Studies. In *Translation Universals. Do they exist?*, pages 15–32. John Benjamins Publishing Company, 2004.

Gideon Toury. *Descriptive Translation Studies and beyond)Revised edition*, volume 100. John Benjamins Publishing, 2012.

Peter D Turney. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44:533–585, 2012.

Kitty van Leuven-Zwart. Translation and original: Similarities and dissimilarities, I. *Target. International Journal of Translation Studies*, 1(2):151–181, 1989.

David Vilar, Jia Xu, Luis Fernando d'Haro, and Hermann Ney. Error analysis of statistical machine translation output. In *Proceedings of LREC*, pages 697–702, 2006.

Jean-Paul Vinay and Jean Darbelnet. *Comparative stylistics of French and English: a methodology for translation*, volume 11. John Benjamins Publishing, 1958/1995.

John White, Theresa O'Connell, and Francis O'Mara. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings of the 1994 Conference of the Association for Machine Translation in the Americas*, pages 193–205, 1994.

Evan James Williams. *Regression Analysis*, volume 14. Wiley, New York, USA, 1959.

Xiaofeng Wu, Hui Yu, and Qun Liu. DCU participation in WMT2013 metrics task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 435–439, 2013.

Xiaofeng Wu, Hui Yu, and Qun Liu. RED: DCU-CASICT Participation in WMT2014 Metrics Task. *Proceedings of ACL-2014*, page 420, 2014.

Muyun Yang, Junguo Zhu, Sheng Li, and Tiejun Zhao. Fusion of Word and Letter Based Metrics for Automatic MT Evaluation. In *IJCAI*, 2013.

Hui Yu, Qingsong Ma, Xiaofeng Wu, and Qun Liu. CASICT-DCU Participation in WMT2015 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 417–421, 2015.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. Re-evaluating Machine Translation Results with Paraphrase Support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 77–84. Association for Computational Linguistics, 2006.

Appendices

Appendix A

RESULTS OF WMT METRICS TASKS

This Appendix contains a detailed presentation of the results for WMT Metrics Tasks datasets.

Metric	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	Average
UPF-Cobalt	.364±.007	.435±.013	.388±.012	.351±.016	.392±.013	.382±.021	.385±.014
Metrics-F	.345±.007	.448±.013	.422±.012	.390±.017	.410±.013	.422±.021	.406±.014
Cobalt-F-comp	.336±.007	.434±.014	.416±.012	.365±.016	.394±.013	.417±.021	.394±.014
Bleu	.278±.007	.357±.014	.270±.013	.274±.016	.320±.013	.246±.024	.291±.014
TER	-.273±.006	-.367±.013	-.286±.012	-.280±.016	-.330±.012	-.241±.021	-.296±.013
Meteor	.329±.007	.426±.013	.379±.013	.341±.016	.385±.013	.342±.023	.367±.014
DPMFcomb	.388±.007	.482±.013	.424±.012	.385±.016	.419±.013	.397±.022	.416±.014
BEER	.342±.007	.416±.014	.376±.013	.332±.016	.373±.013	.370±.023	.368±.014
chrF2	.341±.007	.419±.014	.364±.013	.344±.016	.381±.013	.346±.023	.366±.014
MPEDA	.331±.007	.425±.013	.377±.013	.340±.017	.384±.013	.336±.023	.366±.014
chrF3	.343±.007	.421±.014	.357±.013	.342±.016	.380±.013	.345±.024	.365±.014
chrF1	.323±.007	.410±.014	.377±.013	.339±.017	.376±.013	.345±.023	.362±.015
UoW-ReVal	.261±.007	.377±.014	.334±.013	.313±.018	.314±.014	.340±.022	.323±.015
wordF3	.299±.007	.377±.014	.296±.012	.305±.016	.340±.013	.288±.022	.318±.014
wordF2	.297±.007	.378±.014	.298±.012	.301±.016	.338±.013	.284±.023	.316±.014
wordF1	.290±.007	.370±.014	.295±.013	.293±.017	.334±.013	.276±.024	.310±.015
DTED	.201±.007	.209±.014	.134±.013	.141±.017	.197±.013	.142±.023	.171±.014

Table A.1: Sentence-level evaluation results for WMT16 ranking dataset in terms of Kendall rank correlation coefficient

Metric	fr-en	fi-en	de-en	cs-en	ru-en	Average
UPF-Cobalt	.386±.012	.437±.013	.427±.011	.457±.007	.402±.013	.422±.011
Metrics-F	.413±.012	.450±.011	.457±.013	.502±.011	.410±.013	.447±.011
Cobalt-F-comp	.403±.012	.464±.012	.438±.013	.481±.011	.395±.013	.436±.011
BLEU	.358±.013	.308±.012	.360±.011	.391±.006	.329±.011	.349±.011
Meteor	.380±.012	.406±.011	.422±.010	.439±.008	.386±.011	.407±.010
DPMFCOMB	.395±.012	.445±.012	.482±.009	.495±.007	.418±.013	.447±.011
BEERTREEPEL	.389±.014	.438±.010	.447±.008	.471±.007	.403±.014	.429±.011
RATATOUILLE	.398±.010	.421±.011	.441±.010	.472±.007	.393±.013	.425±.010
BEER	.393±.012	.422±.012	.438±.010	.457±.008	.396±.014	.421±.011
CHRF	.383±.011	.417±.012	.424±.010	.446±.008	.384±.014	.411±.011
CHRF3	.383±.013	.397±.011	.421±.010	.449±.008	.386±.013	.407±.011
METEOR-WSD	.375±.012	.406±.010	.420±.011	.438±.008	.387±.012	.405±.010
DPMF	.368±.012	.411±.011	.418±.011	.436±.008	.378±.011	.402±.011
LEBLEU-OPTIMIZED	.376±.013	.391±.010	.399±.010	.438±.008	.374±.012	.396±.011
LEBLEU-DEFAULT	.373±.013	.383±.011	.402±.009	.436±.007	.376±.011	.394±.010
VERTA-EQ	.388±.012	.369±.013	.410±.011	.447±.007	.346±.013	.392±.011
VERTA-70ADEQ30FLU	.374±.012	.365±.014	.418±.011	.438±.007	.344±.013	.388±.011
VERTA-W	.383±.010	.344±.014	.416±.010	.445±.007	.345±.013	.387±.011
DREEM	.362±.012	.340±.010	.368±.011	.423±.007	.348±.013	.368±.011
UOW-LSTM	.332±.011	.376±.012	.375±.011	.385±.008	.356±.010	.365±.011
TOTAL-BS	.332±.013	.319±.013	.333±.010	.381±.007	.321±.011	.337±.011

Table A.2: Sentence-level evaluation results for WMT15 ranking dataset in terms of Kendall rank correlation coefficient

Metric	fr-en	de-en	hi-en	cs-en	ru-en	Avg
UPF-Cobalt	.402±.012	.348±.013	.439±.013	.315±.016	.330±.010	.367±.013
DISCOTK-PARTY-TUNED	.433±.012	.380±.013	.434±.013	.328±.015	.355±.011	.386±.013
BEER	.417±.013	.337±.014	.438±.013	.284±.016	.333±.011	.362±.013
REDCOMBSENT	.406±.012	.338±.014	.417±.013	.284±.015	.336±.011	.356±.013
REDCOMBSYSSENT	.408±.012	.338±.014	.416±.013	.282±.014	.336±.011	.356±.013
METEOR	.406±.012	.334±.014	.420±.013	.282±.015	.329±.010	.354±.013
REDSYSSENT	.404±.012	.338±.014	.386±.014	.283±.015	.321±.010	.346±.013
REDSENT	.403±.012	.336±.014	.383±.014	.283±.015	.323±.011	.345±.013
UPC-IPA	.412±.012	.340±.014	.368±.014	.274±.015	.316±.011	.342±.013
UPC-STOUT	.403±.012	.345±.014	.352±.014	.275±.015	.317±.011	.338±.013
VERTA-W	.399±.013	.321±.015	.386±.014	.263±.015	.315±.011	.337±.014
VERTA-EQ	.407±.013	.315±.014	.384±.013	.263±.015	.312±.011	.336±.013
DISCOTK-PARTY	.395±.013	.334±.014	.362±.013	.264±.016	.305±.011	.332±.013
AMBER	.367±.013	.313±.014	.362±.013	.246±.016	.294±.011	.316±.013
BLEUNRC	.382±.013	.272±.014	.322±.014	.226±.016	.269±.011	.294±.013
SENTBLEU	.378±.013	.271±.014	.300±.013	.213±.016	.263±.011	.285±.013
APAC	.364±.012	.271±.014	.288±.014	.198±.016	.276±.011	.279±.013
DISCOTK-LIGHT	.311±.014	.224±.015	.238±.013	.187±.016	.209±.011	.234±.014
DISCOTK-LIGHT-KOOL	.005±.001	.001±.000	.000±.000	.002±.001	.001±.000	.002±.001

Table A.3: Sentence-level evaluation results for WMT14 ranking dataset in terms of Kendall rank correlation coefficient

Metric	cs-en	de-en	es-en	fr-en	ru-en	Average
UPF-Cobalt	.253±.006	.300±.005	.319±.007	.267±.007	.227±.005	.273±.006
Meteor	.256±.006	.280±.005	.304±.006	.250±.006	.229±.004	.264±.006
BLEU	.187±.006	.203±.005	.240±.007	.204±.007	.154±.005	.197±.006
Depref-align	.218±.006	.252±.005	.289±.007	.239±.006	.190±.005	.238±.006
Depref-exact	.217±.006	.247±.005	.282±.007	.238±.006	.185±.005	.234±.006
SIMPBLEU_RECALL	.192±.006	.230±.005	.271±.006	.209±.005	.171±.004	.215±.005
SIMPBLEU_PREC	.196±.006	.219±.005	.261±.007	.215±.007	.163±.004	.211±.006

Table A.4: Sentence-level evaluation results for WMT13 ranking dataset in terms of Kendall rank correlation coefficient

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	Average
UPF-Cobalt	.364±.007*	.435±.013	.388±.012	.351±.016	.392±.013*	.382±.021	.385±.014*
meteor aligner	.318±.007	.419±.015	.362±.012	.335±.017	.372±.013	.338±.022	.357±.014
penalty off	.342±.007	.401±.014	.394±.012*	.342±.016	.373±.013	.374±.021	.371±.014
weights off	.357±.007*	.433±.014*	.382±.013*	.355±.016*	.388±.013*	.375±.023*	.382±.014*
equivalence off	.357±.007	.436±.014	.389±.013	.348±.016	.389±.013	.382±.022*	.384±.014
embeddings off	.349±.007	.437±.014*	.385±.012	.359±.016*	.387±.013	.354±.021	.378±.014

Table A.5: Ablations tests results for WMT16 ranking task

	fr-en	fi-en	de-en	cs-en	ru-en	Average
UPF-Cobalt	.384±.012	.436±.011	.429±.010	.459±.007	.407±.011	.423±.010
equivalence off	.383±.012	.433±.011	.430±.010	.461±.007	.404±.011	.422±.010
weights off	.380±.012	.430±.011	.423±.010	.461±.007	.402±.011	.419±.010
penalty off	.378±.012	.425±.012	.417±.009	.466±.007	.404±.011	.418±.010
embeddings off	.388±.013	**.414±.012**	.420±.010	**.422±.007**	**.391±.010**	.407±.010
aligner off	**.369±.013**	**.392±.011**	**.395±.011**	**.413±.008**	**.368±.010**	**.387±.011**

Table A.6: Ablations test results for WMT15 ranking task

	fr-en	de-en	hi-en	cs-en	ru-en	Avg
UPF-Cobalt	.402±.012	.348±.013	.439±.013	.315±.016	.330±.010	.367±.013
cobalt	.398±.013	.346±.014	.442±.012	.309±.014	.328±.011	.365±.013
equivalence off	.399±.013	.343±.014	.445±.013	.301±.015	.324±.011	.362±.013
embeddings off	.397±.013	.342±.014	.435±.013	.291±.015	.330±.010	.359±.013
weights off	.395±.013	.340±.014	.435±.013	.295±.015	.322±.011	.358±.013
penalty off	.378±.013	.341±.014	.437±.013	.291±.014	.322±.011	.354±.013
aligner off	.386±.013	.317±.014	.401±.013	.282±.014	.304±.011	.338±.013

Table A.7: Ablations test results for WMT14 ranking task

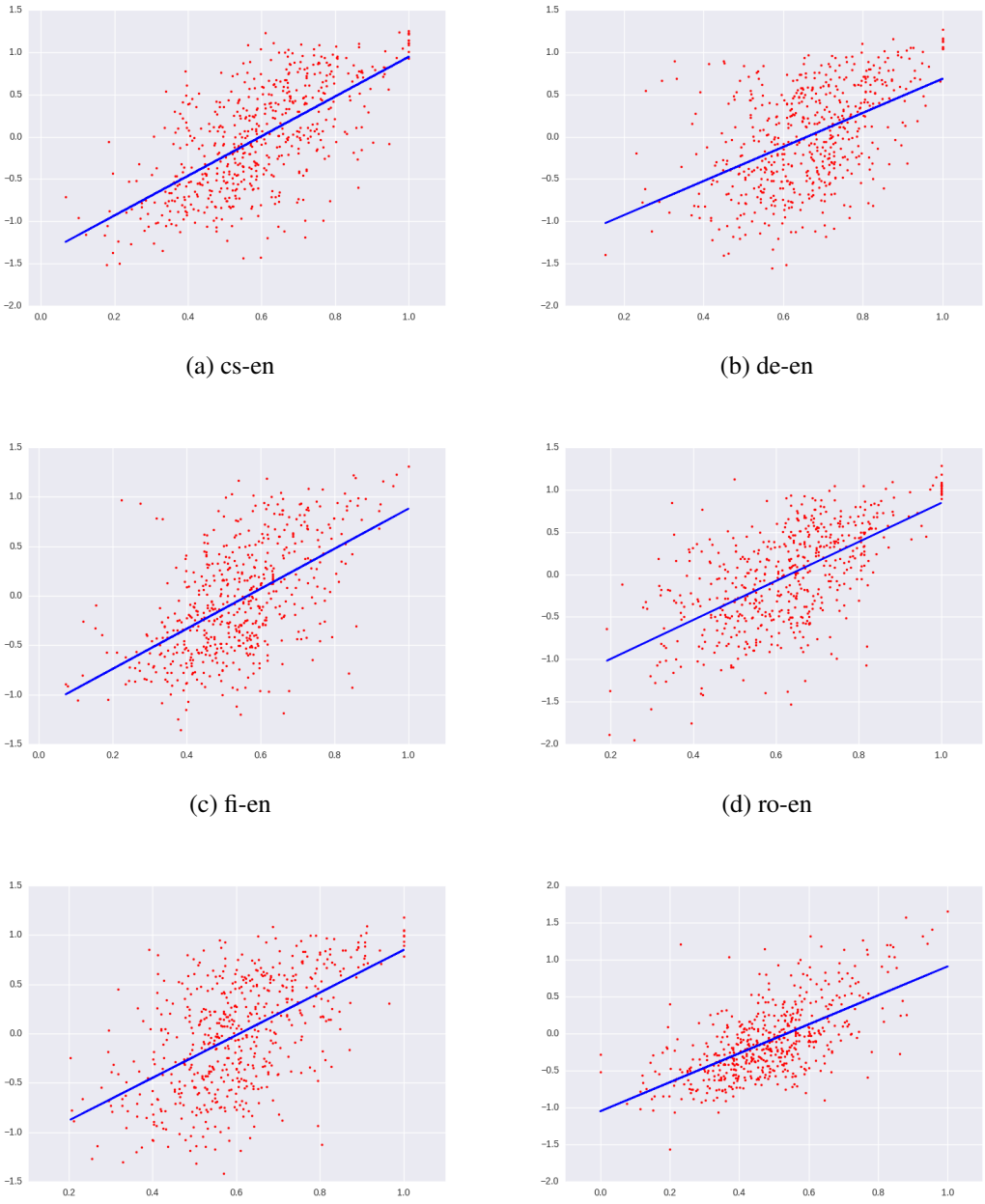
	cs-en	de-en	fi-en	ro-en	ru-en	tr-en
UPF-Cobalt	0.652	0.490	0.550	0.616	0.556	0.626
BLEU	$0.568^{\dagger}$	0.447	$0.433^{\dagger}$	$0.499^{\dagger}$	$0.470^{\dagger}$	$0.538^{\dagger}$
Meteor	0.645	0.517	0.540	0.587	0.548	0.618
TER	$-0.578^{\dagger}$	$-0.468^{\dagger}$	$-0.411^{\dagger}$	$-0.441^{\dagger}$	$-0.459^{\dagger}$	$-0.491^{\dagger}$
MPEDA	0.644	0.513	0.538	0.587	0.545	0.616
DPMFcomb	$0.713^{\dagger}$	$0.598^{\dagger}$	$0.584^{\dagger}$	0.627	$0.615^{\dagger}$	$0.663^{\dagger}$
chrF2	0.658	0.469	$0.457^{\dagger}$	$0.581^{\dagger}$	0.534	$0.556^{\dagger}$
BEER	0.661	0.471	$0.462^{\dagger}$	$0.551^{\dagger}$	0.533	$0.545^{\dagger}$

Table A.8: Pearson correlation with WMT16 direct assessments

	L1	L2	L3	L4
upf-cobalt	0.150	0.122	0.170	0.403
cobalt-f-comp	$0.092^{\dagger}$	$0.160^{\dagger}$	$0.216^{\dagger}$	$0.469^{\dagger}$
metrics-f	0.127	$0.172^{\dagger}$	$0.199^{\dagger}$	$0.480^{\dagger}$
BLEU	$0.034^{\dagger}$	0.084	0.134	$0.456^{\dagger}$
Meteor	$0.198^{\dagger}$	0.151	0.163	$0.514^{\dagger}$
TER	$-0.040^{\dagger}$	$-0.063^{\dagger}$	$-0.179^{\dagger}$	$-0.388^{\dagger}$
MPEDA	$0.200^{\dagger}$	0.150	0.166	$0.512^{\dagger}$
DPMFcomb	$0.204^{\dagger}$	0.146	0.193	$0.443^{\dagger}$
chrF2	$0.218^{\dagger}$	0.119	0.139	0.375
BEER	$0.228^{\dagger}$	0.119	0.143	0.384

Table A.9: Pearson correlation for WMT16 direct assessment dataset with different quality levels (L1-L4)

Figures A.1 - A.10 show the scatter plots for the correlation between WMT16 direct assessments (DA) and the scores of our metrics, the usual benchmark metrics BLEU, Meteor and TER, as well as the metrics that participated in WMT16 Metrics Task.



(a) cs-en

(b) de-en

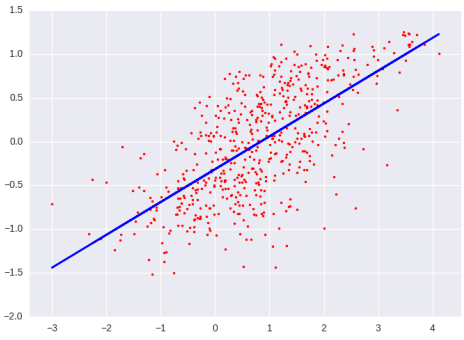
(c) fi-en

(d) ro-en

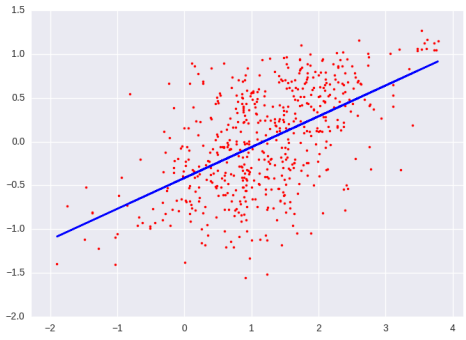
(e) ru-en

(f) tr-en

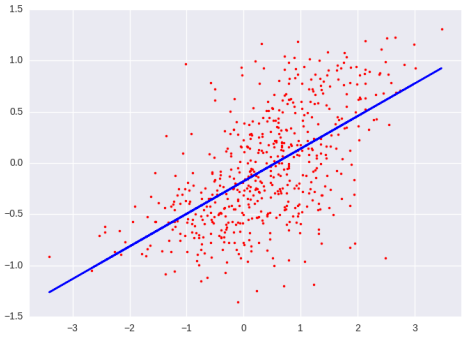
Figure A.1: Scatter plots for UPF-Cobalt scores and DA human judgments for WMT16 DA dataset



(a) cs-en

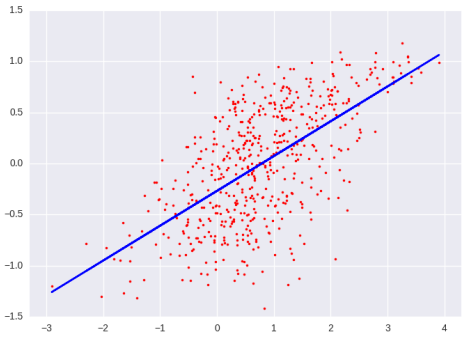


(b) de-en

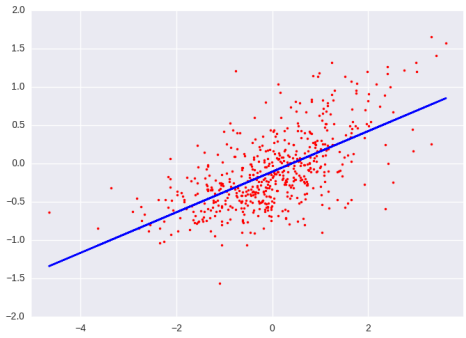


(c) fi-en

(d) ro-en

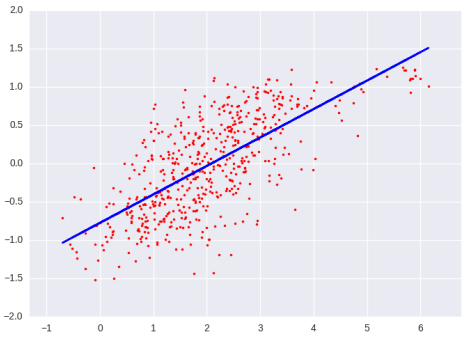


(e) ru-en

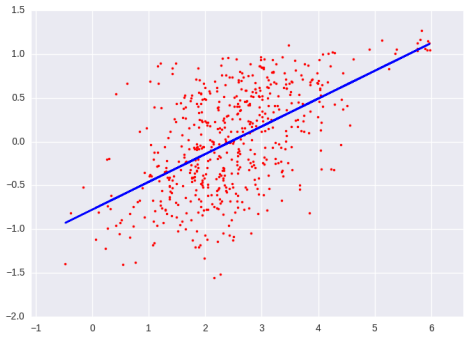


(f) tr-en

Figure A.2: Scatter plots for Cobalt-Fcomp scores and DA human judgments for WMT16 DA dataset

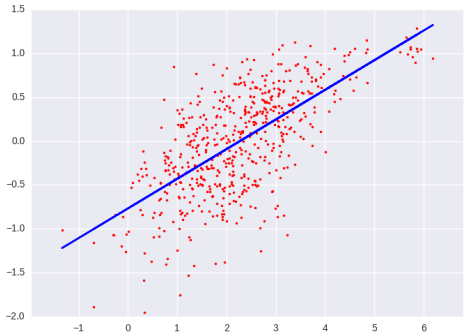


(a) cs-en

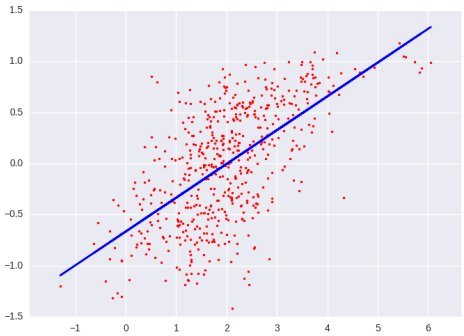


(b) de-en

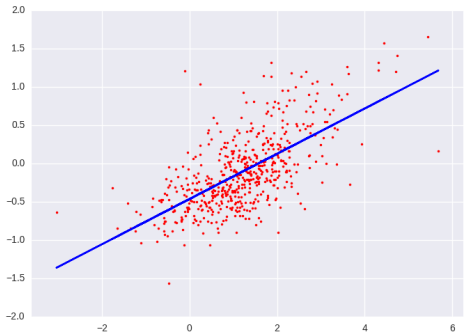
(c) fi-en



(d) ro-en

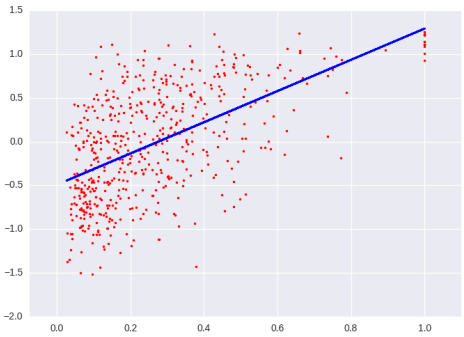


(e) ru-en

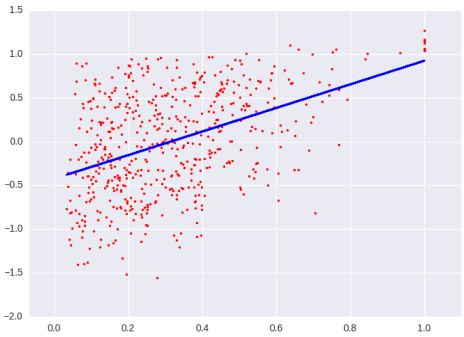


(f) tr-en

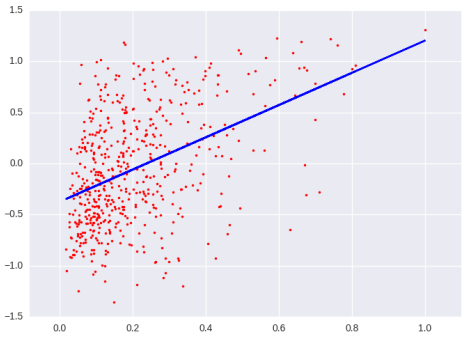
Figure A.3: Scatter plots for Metrics-F scores and DA human judgments for WMT16 DA dataset



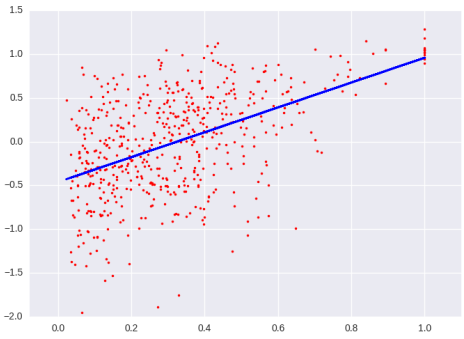
(a) cs-en



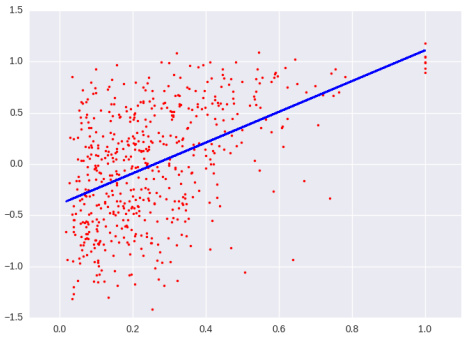
(b) de-en



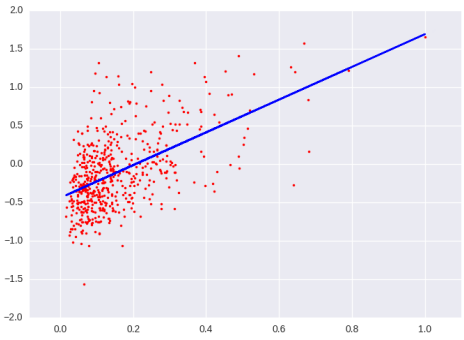
(c) fi-en



(d) ro-en

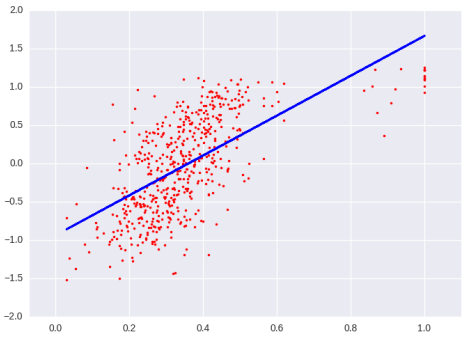


(e) ru-en

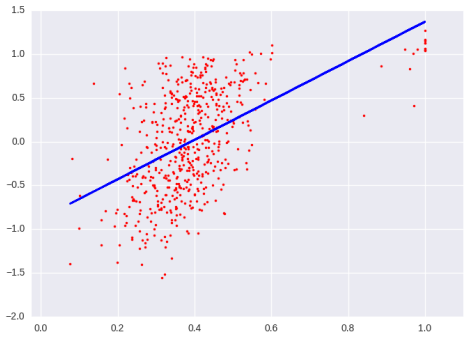


(f) tr-en

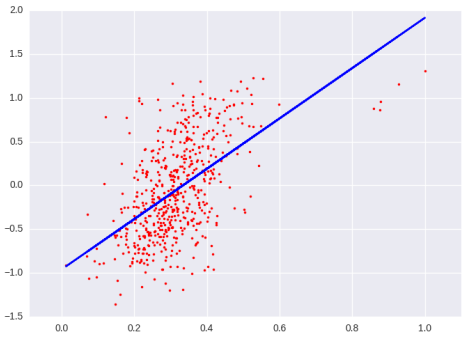
Figure A.4: Scatter plots for BLEU scores and DA human judgments for WMT16 DA dataset



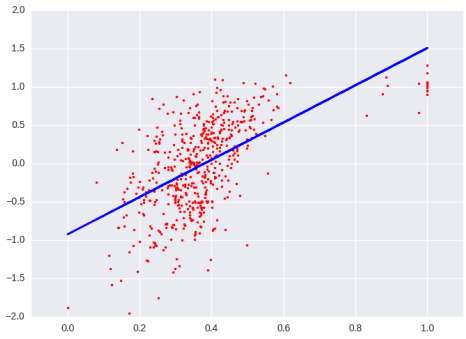
(a) cs-en



(b) de-en

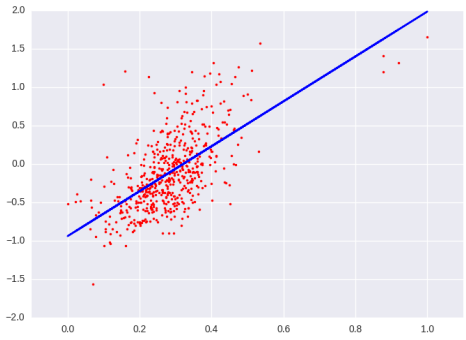


(c) fi-en



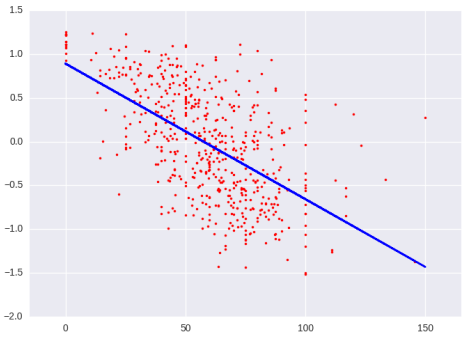
(d) ro-en

(e) ru-en



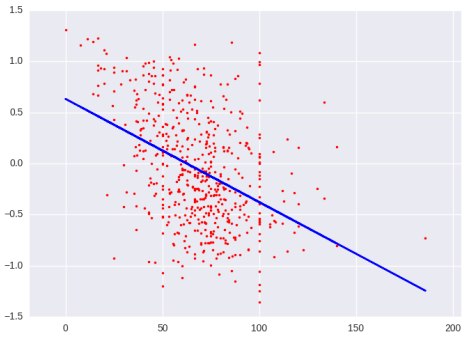
(f) tr-en

Figure A.5: Scatter plots for Meteor scores and DA human judgments for WMT16 DA dataset

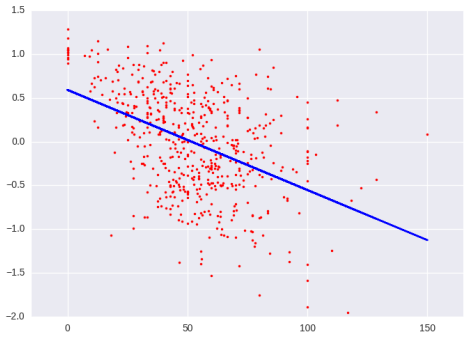


(a) cs-en

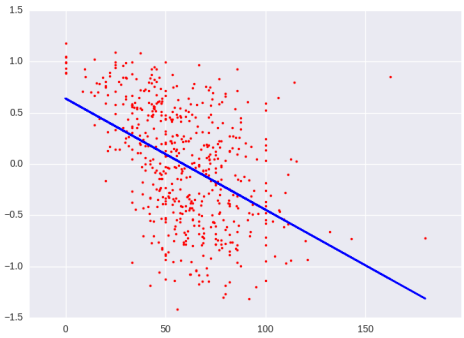
(b) de-en



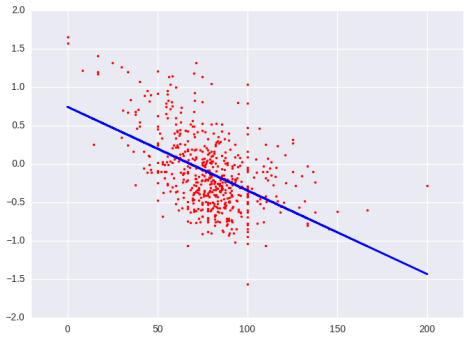
(c) fi-en



(d) ro-en

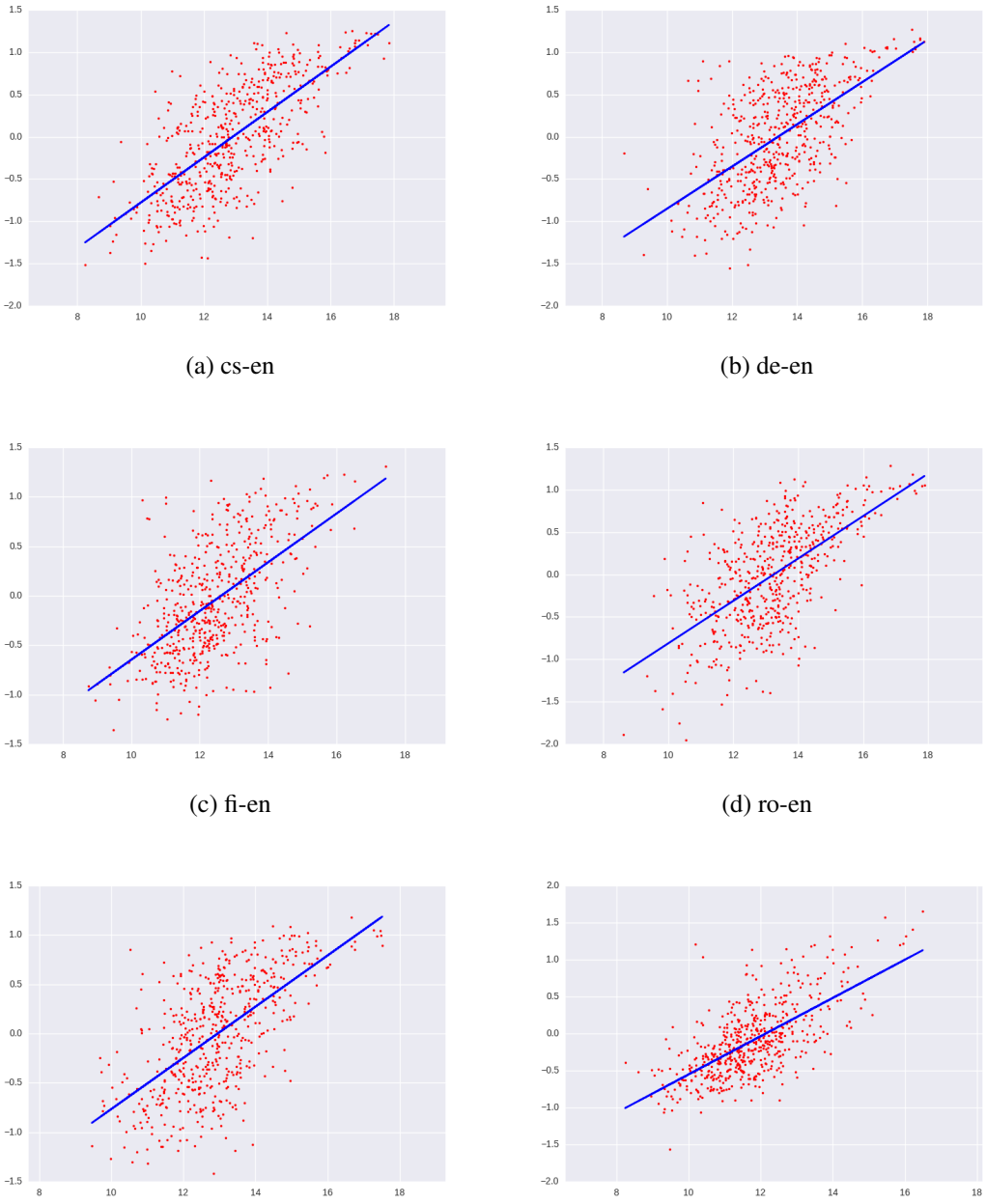


(e) ru-en



(f) tr-en

Figure A.6: Scatter plots for TER scores and DA human judgments for WMT16 DA dataset



(a) cs-en

(b) de-en

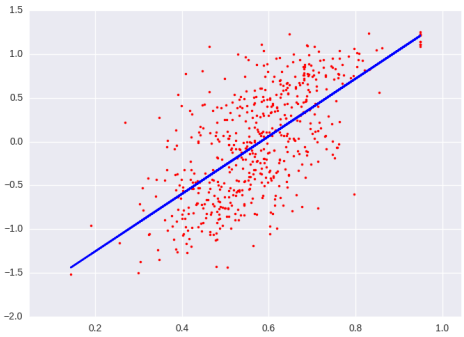
(c) fi-en

(d) ro-en

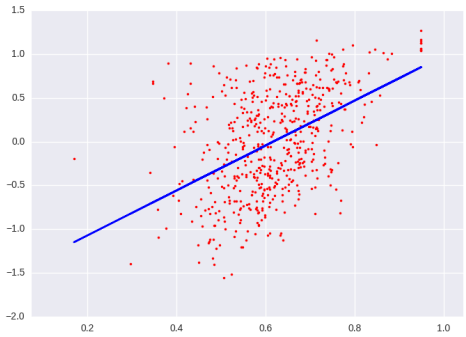
(e) ru-en

(f) tr-en

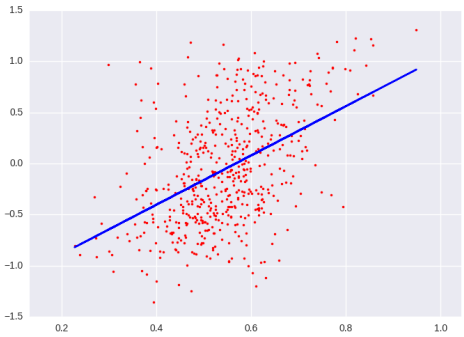
Figure A.7: Scatter plots for DPMFcomb scores and DA human judgments for WMT16 DA dataset



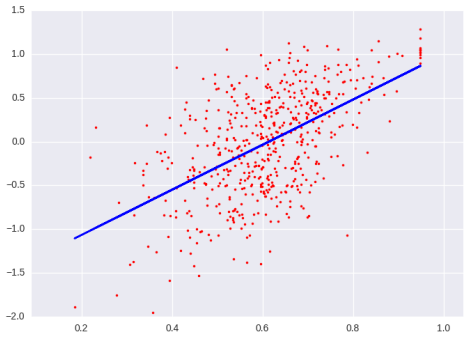
(a) cs-en



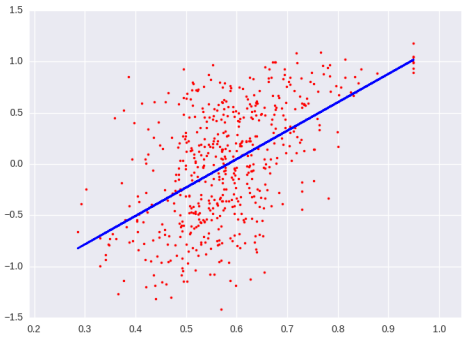
(b) de-en



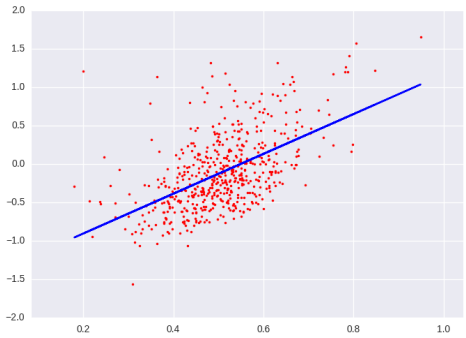
(c) fi-en



(d) ro-en

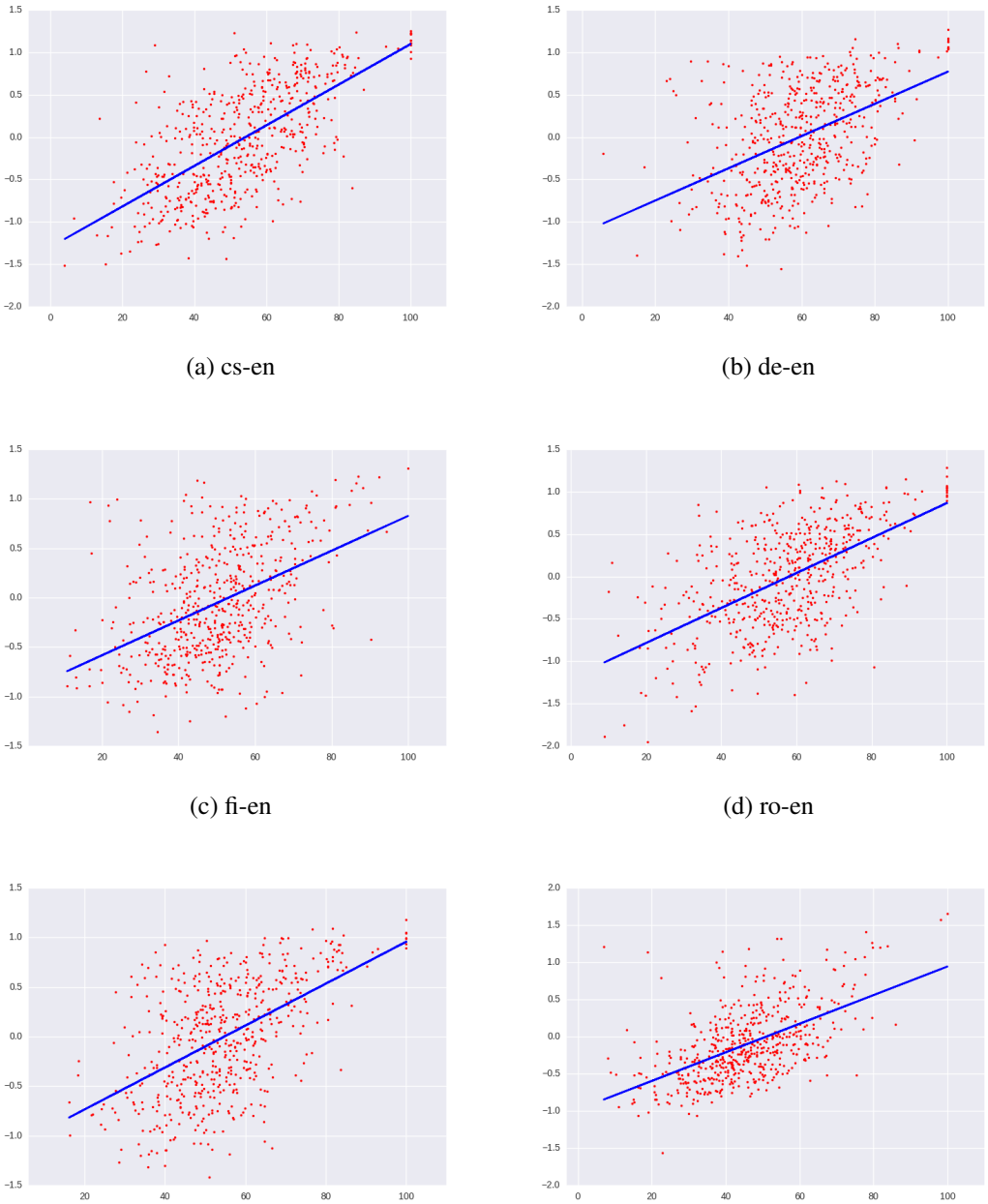


(e) ru-en



(f) tr-en

Figure A.8: Scatter plots for BEER scores and DA human judgments for WMT16 DA dataset

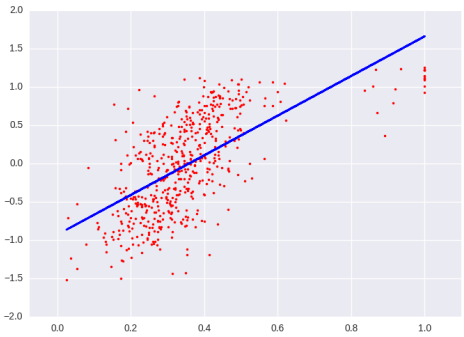


(a) cs-en (b) de-en

(c) fi-en (d) ro-en

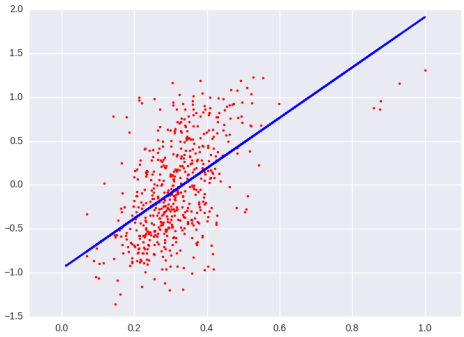
(e) ru-en (f) tr-en

Figure A.9: Scatter plots for ChrF2 scores and DA human judgments for WMT16 DA dataset

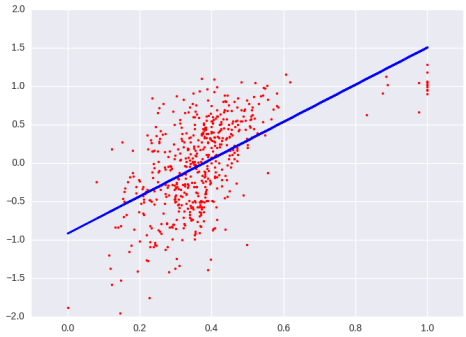


(a) cs-en

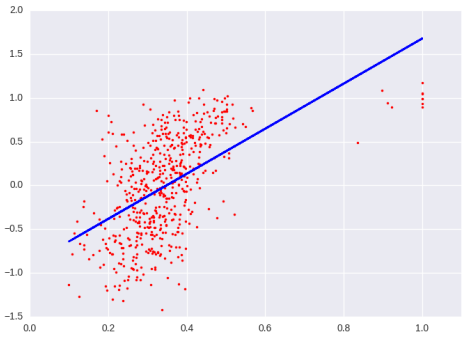
(b) de-en



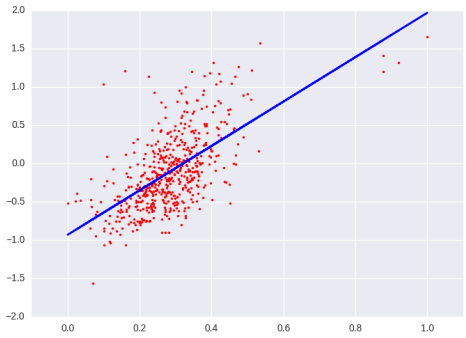
(c) fi-en



(d) ro-en



(e) ru-en



(f) tr-en

Figure A.10: Scatter plots for MPEDA scores and DA human judgments for WMT16 DA dataset

Appendix B

COBALT-F FEATURES

This Appendix contains a description of adequacy-oriented features (Table B.1) and fluency-oriented features (Table B.2) used in our Cobalt-F metric presented in Chapter 5.

avg_pen_tgt	Average context penalty with respect to the candidate translation
avg_pen_exact_tgt	Average context penalty with respect to the candidate translation, for words that constitute an exact match
avg_pen_exact_ref	Average context penalty with respect to the reference translation, for words that constitute an exact match
avg_pen_non_exact_tgt	Average context penalty with respect to the candidate translation, for words that do not constitute an exact match
avg_pen_non_exact_ref	Average context penalty with respect to the reference translation, for words that do not constitute an exact match
avg_pen_ref	Average context penalty with respect to the reference translation
count_aligned_content	Number of aligned content words
count_aligned_function	Number of aligned function words
count_aligned	Number of aligned words
count_aligned_content_tgt	Number of aligned content words in the candidate translation
count_aligned_content_ref	Number of aligned content words in the reference translation

count_aligned_function_tgt	Number of aligned function words in the candidate translation
count_aligned_function_ref	Number of aligned function words in the reference translation
count_non_aligned_tgt	Number of non-aligned words in the candidate translation
count_non_aligned_content_tgt	Number of non-aligned content words in the candidate translation
count_non_aligned_function_tgt	Number of non-aligned function words in the candidate translation
count_non_aligned_ref	Number of non-aligned words in the reference translation
count_pen	Number of words with context penalty
count_words_tgt	Total number of words in the candidate translation
count_words_ref	Total number of words in the reference translation
lengths_ratio	Ratio of candidate and reference lengths
max_pen_tgt	Maximum context penalty with respect to the candidate translation
max_pen_ref	Maximum context penalty with respect to the reference translation
min_pen_tgt	Minimum context penalty with respect to the candidate translation
min_pen_ref	Minimum context penalty with respect to the reference translation
prop_aligned_tgt	Proportion of aligned words in the candidate translation
prop_aligned_content_tgt	Proportion of aligned content words in the candidate translation
prop_aligned_content_ref	Proportion of aligned content words in the reference translation
prop_aligned_function_tgt	Proportion of aligned function words in the candidate translation
prop_aligned_function_ref	Proportion of aligned function words in the reference translation
prop_aligned_ref	Proportion of aligned words in the reference translation
prop_lex_distrib	Proportion of words aligned based on distributional similarity

prop_lex_exact	Proportion of words aligned based on exact, stem or lemma match
prop_lex_para	Proportion of words aligned based on paraphrase match
prop_lex_syn	Proportion of words aligned based on synonym match
prop_non_aligned_tgt	Proportion of non-aligned words in the candidate translation
prop_non_aligned_content_tgt	Proportion of non-aligned content words in the candidate translation
prop_non_aligned_function_tgt	Proportion of non-aligned function words in the candidate translation
prop_non_aligned_ref	Proportion of non-aligned words in the reference translation
prop_pen_exact_tgt	Proportion of candidate words with context penalty out of the total number of words aligned based on exact match
prop_pen_exact_ref	Proportion of reference words with context penalty out of the total number of words aligned based on exact match
prop_pen_high	Proportion of words with high context penalty out of the total number of aligned words
prop_pen_non_exact_tgt	Proportion of candidate words with context penalty out of the total number of words aligned based on non-exact match
prop_pen_non_exact_ref	Proportion of reference words with context penalty out of the total number of words aligned based on non-exact match
prop_pen	Proportion of words with context penalty out of the total number of aligned words
prop_pos_coarse	Proportion of aligned words that belong to the same grammatical category
prop_pos_diff	Proportion of aligned words that belong to different grammatical categories
prop_pos_exact	Proportion of aligned words with exactly matching POS tags

Table B.1: Adequacy-oriented Features

count_content_tgt	Number of content words in the candidate translation
count_function_tgt	Number of function words in the candidate translation
count_words_tgt	Number of words in the candidate translation
lang_mod_perplex	LM perplexity of the candidate translation
lang_mod_perplex2	LM perplexity of the candidate translation without the end-of-sentence marker
lang_mod_prob	LM probability of the candidate translation
pos_lang_mod_perplex_srilm	POS LM perplexity of the candidate translation
pos_lang_mod_prob_srilm	POS LM probability of the candidate translation
prop_adjectives	proportion of adjectives in the candidate translation
prop_nouns	proportion of nouns in the candidate translation
prop_verbs	proportion of verbs in the candidate translation
prop_verbs_flex	proportion of inflected verbs in the candidate translation
count_adjectives	number of adjectives in the candidate translation
count_nouns	number of nouns in the candidate translation
count_verbs	number of verbs in the candidate translation
count_verbs_flex	number of inflected verbs in the candidate translation
backoff_avg	Average backoff behaviour value
backoff_back_avg	Average backoff behaviour value using backward LM
backoff_back_max	Maximum backoff behaviour value using backward LM
backoff_back_median	Median backoff behaviour value using backward LM

backoff_back_min	Minimum backoff behaviour value using backward LM
backoff_max	Maximum backoff behaviour value
backoff_median	Median backoff behaviour value
backoff_min	Minimum backoff behaviour value
count_oov	Number of out-of-vocabulary words
prop_oov	Proportion of out-of-vocabulary words
backoff_back_non_aligned_avg	Average backoff behaviour value using backward LM for non-aligned words
backoff_back_non_aligned_max	Maximum backoff behaviour value using backward LM for non-aligned words
backoff_back_non_aligned_median	Median backoff behaviour value using backward LM for non-aligned words
backoff_back_non_aligned_min	Minimum backoff behaviour value using backward LM for non-aligned words
backoff_non_aligned_avg	Average backoff behaviour value for non-aligned words
backoff_non_aligned_max	Maximum backoff behaviour value for non-aligned words
backoff_non_aligned_median	Median backoff behaviour value for non-aligned words
backoff_non_aligned_min	Minimum backoff behaviour value for non-aligned words
count_backoff_low_non_aligned	Number of non-aligned words with low backoff behaviour value (< 5)
count_non_aligned_oov	Number of non-aligned out-of-vocabulary words
prop_backoff_low_non_aligned	Proportion of non-aligned words with low backoff behaviour value (< 5)
prop_non_aligned_oov	Proportion of non-aligned out-of-vocabulary words
pos_back_direct_non_aligned_avg	Average backoff behaviour value using backward POS LM for non-aligned words
pos_back_direct_non_aligned_max	Maximum backoff behaviour value using backward POS LM for non-aligned words
pos_back_direct_non_aligned_median	Median backoff behaviour value using backward POS LM for non-aligned words
pos_back_direct_non_aligned_min	Minimum backoff behaviour value using backward POS LM for non-aligned words
prop_adjectives_non_aligned_tgt	Proportion of non-aligned adjectives in the candidate translation

prop_adjectives_non_aligned_ref	Proportion of non-aligned adjectives in the reference translation
prop_nouns_non_aligned_tgt	Proportion of non-aligned nouns in the candidate translation
prop_nouns_non_aligned_ref	Proportion of non-aligned nouns in the reference translation
prop_verbs_flex_non_aligned_tgt	Proportion of non-aligned inflected verbs in the candidate translation
prop_verbs_flex_non_aligned_ref	Proportion of non-aligned inflected verbs in the reference translation
prop_verbs_non_aligned_tgt	Proportion of non-aligned verbs in the candidate translation
prop_verbs_non_aligned_ref	Proportion of non-aligned verbs in the reference translation
pos_backoff_direct_avg	Average backoff behaviour value using POS LM for non-aligned words
pos_backoff_direct_max	Maximum backoff behaviour value using POS LM for non-aligned words
pos_backoff_direct_median	Median backoff behaviour value using POS LM for non-aligned words
pos_backoff_direct_min	Minimum backoff behaviour value using POS LM for non-aligned words

Table B.2: Fluency-oriented Features