Chapter 4

Multilingual FEMsum

This chapter shows how combining different components of our general architecture, presented in Chapter 3, leads to different multilingual summarization approaches. All of them have been developed in the framework of the Spanish Research Department funded projects HERMES and ALIADO.

The project *Hemerotecas Electrónicas, Recuperación Multilingüe y Extracción Semántica (HERMES)*[1] deals with requests for information contained in a multilingual digital news archive. Applications developed within HERMES facilitate the retrieval of multilingual textual information and automatic summaries are produced for any document related to a user query.

The project *Tecnologías del habla y el lenguaje para un asistente personal (ALIADO)*[2] undertakes the developing of spoken and written language technologies for the design of personal assistants in a multilingual environment. Special attention is paid to the design of the oral interface. One of the examples of language centered help in the project is the "Question & Answering" facility, where the answer may be a summary.

This chapter is organized in three sections. In Section 4.1 approaches that exploit discourse textual properties are evaluated in summarizing Spanish and Catalan documents. Section 4.2 presents a mixed FEMsum instantiation originally designed to produce English headlines for the DUC 2003 Task1. A ML approach is applied for detecting the most relevant sentence, which is then compressed by manual rules to obtain a grammatical headline. Finally, Section 4.3 concludes the chapter.

[1] *Electronic Newspaper Libraries, Multilingual Retrieval and Semantic Extraction*
`http://nlp.uned.es/hermes/` (TIC2000-0335-C03-02)

[2] *Language and Speech Technologies for a Personal Assistant*
`http://gps-tsc.upc.es/veu/aliado/` (TIC2002-04447)

4.1 Agency news multilingual Single Document Summarization

This section presents and analyzes *LCsum*, a FEMsum approach that exploits textual properties. *LCsum* has been instantiated with different FEMsum components to study the influence of using linguistic information in well-written SDS. In particular, the role of cohesion and coherence in summary content selection has been analyzed.

The SDS task studied in this section consists in summarizing a piece of news by producing extract based summaries of about 10% of the original document length. The input document could be written in Catalan, Spanish or English. However, we have focused our efforts on the development of resources for Catalan and Spanish.

For this task, several *LCsum* settings have been evaluated using the framework described in Section 4.1.1. The analysis of the performance of *LCsum* when summarizing agency news documents in Spanish and Catalan is presented in Section 4.1.2.

4.1.1 Multilingual evaluation framework

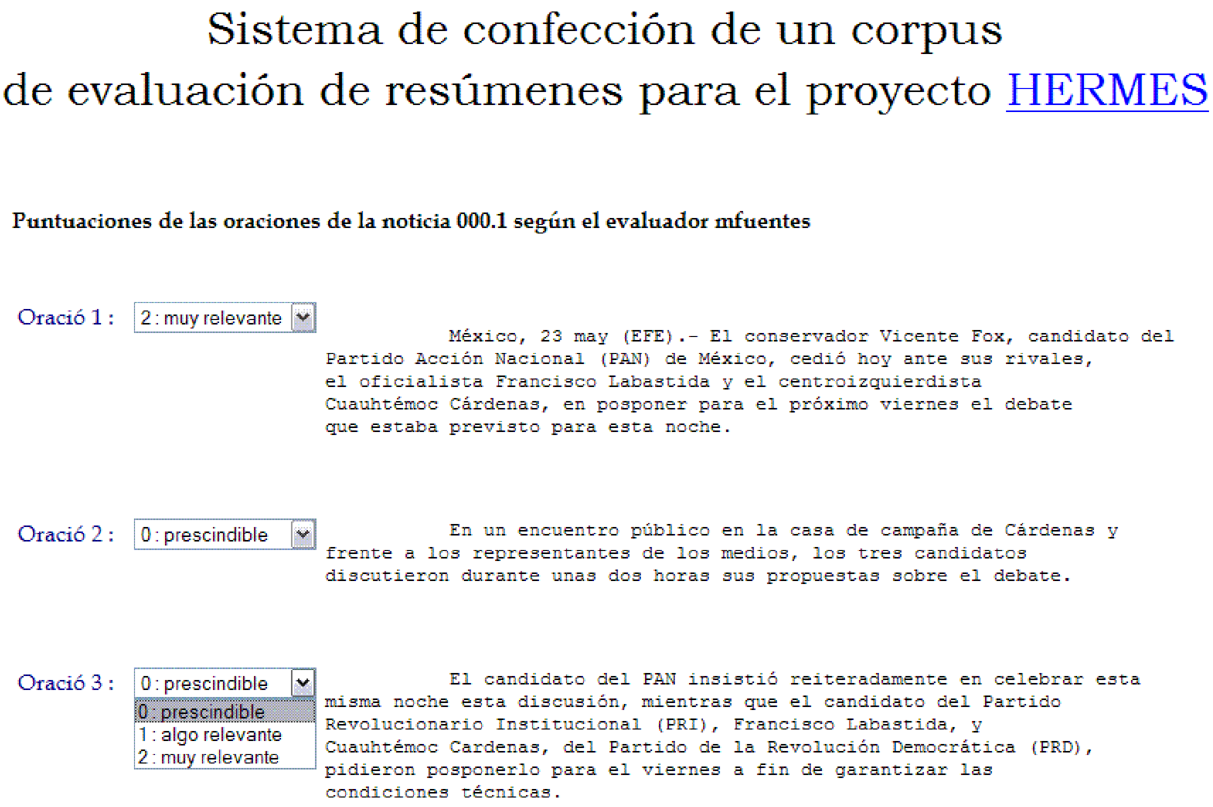
As said in previous chapters, to automatically evaluate summaries a distinction is made between extrinsic, task-based evaluation and intrinsic evaluation, taking into account the quality of the summary by itself or by a comparison with a *gold standard.* A *gold standard* is the ideal summary that the system is intended to produce. In the context of the HERMES project a *gold standard* test corpus was manually created to evaluate Spanish SDS approaches.

To avoid the variability of human generated abstracts, an extract-based *gold standard* was created from a corpus of 120 news agency stories (reduced to 111 after removing news with only one paragraph). The TUs to be extracted were paragraph-sized, since they were found to be natural meaning units.

This corpus contains documents of various topics, including economy, finance, politics, science, education, sport, meteorology, health, and society. Stories range from 2 to 28 sentences and from 28 to 734 words in length, with an average length of 275 words per story. The group of news was randomly selected from a corpus provided by EFE, a Spanish news agency.

The *gold standard* was created by 31 human evaluators. Each subject summarized a set of articles numbered from 1 to 77. The objective was to have at less 5 different summaries made for each article.

The human summary was made via web with the interface in Figure 4.1. Each piece of news was presented in turn to the evaluator, segmented at paragraph level. Paragraph TUs were numbered so that they could be easily referenced. In order to deal with different compression degrees, human evaluators were asked to assign a score to each of the TUs of the article. Three possible scores, [0,1,2], were used to mark the relevance of the TU in the whole article. In the



Sistema de confección de un corpus
de evaluación de resúmenes para el proyecto HERMES

Puntuaciones de las oraciones de la noticia 000.1 según el evaluador mfuentes

Oració 1 : 2 : muy relevante

México, 23 may (EFE).- El conservador Vicente Fox, candidato del Partido Acción Nacional (PAN) de México, cedió hoy ante sus rivales, el oficialista Francisco Labastida y el centroizquierdista Cuauhtémoc Cárdenas, en posponer para el próximo viernes el debate que estaba previsto para esta noche.

Oració 2 : 0 : prescindible

En un encuentro público en la casa de campaña de Cárdenas y frente a los representantes de los medios, los tres candidatos discutieron durante unas dos horas sus propuestas sobre el debate.

Oració 3 : 0 : prescindible

El candidato del PAN insistió reiteradamente en celebrar esta misma noche esta discusión, mientras que el candidato del Partido Revolucionario Institucional (PRI), Francisco Labastida, y Cuauhtémoc Cardenas, del Partido de la Revolución Democrática (PRD), pidieron posponerlo para el viernes a fin de garantizar las condiciones técnicas.

Figure 4.1: Web interface to create the Spanish summarization corpus.

instructions, the term relevance was loosely defined. Essentially, the meaning of relevance 2 is "*This TU would occur in my summary*" and the meaning of relevance 0 is "*This TU wouldn't occur in my summary*". Each evaluator was asked to provide a list of keywords for each document as well.

Two different gold standards were obtained from these scores, one containing summaries coming as close as possible to the 10% of the length of the original text (resulting on an average 19% compression) and the other containing the best summaries. We defined the best summary as a group of TUs with more than a half of the maximum possible score. This resulted on an average of 31% of the length of the original text (29% compression).

For the experiments reported in this section, the first set of summaries was taken as *gold standard*. In the evaluation process we used the MEADeval[3] evaluation software developed within the MEAD project. From this package the usual *Precision* and *Recall* measures have

[3] http://tangra.si.umich.edu/clair/meadeval/

been used. The *Cosine*[4] measure was also useful to evaluate the content of those summary TUs that are not in the gold standard.

As a comparison ground to evaluate approaches to summarizing Spanish agency news, two baseline systems were used. The first was the *lead* method, i.e. extracting a number of sentences, starting from the first one, until the desired length, or compression rate, is achieved. The second baseline was provided by the *SweSum* (Dalianis 2000) system. Summaries with *SweSum* were produced with the default parameters of the system.

Because of the cost of building an evaluation corpus, in our experiments in Catalan both documents and summaries were automatically translated from the Spanish corpus to Catalan using the interNOSTRUM (Canals-Marote et al. 2001) Machine Translation (MT) system.

4.1.2 Exploiting Textual Properties

The approaches presented in this section exploit textual properties. Traditionally, two main properties have been distinguished in the discursive structure of a source text: cohesion and coherence. As defined by (Halliday and Hasan 1976), **cohesion** tries to account for relationships among the elements of a text. Four broad categories of cohesion are identified: *reference*, *ellipsis*, *conjunction*, and *lexical cohesion*. On the other hand, **coherence** is represented in terms of relations between text segments, such as *elaboration*, *cause* or *explanation*. Thus, coherence defines the macro-level semantic structure of a connected discourse, while cohesion creates connectedness in a non-structural manner.

The rest of the section is structured as follows: First, we present a *LCsum* prototype that takes into account the text lexical cohesion. This first prototype is evaluated on summarizing Spanish (Fuentes and Rodríguez 2002) and Catalan (Fuentes et al. 2004) documents. After that, the way of adding coherence textual properties is studied when producing summaries from Spanish documents (Alonso and Fuentes 2002; Alonso and Fuentes 2003).

Using cohesive textual properties to summarize Spanish documents

The *LCsum* starting point is an extractive summarization system that exploits the cohesive properties of text by building and ranking lexical chains. This first SDS FEMsum instantiation is based on (Barzilay 1997)'s work. The main textual property to be exploited was lexical cohesion, a concept based on the work by (Halliday and Hasan 1976), where they checked the frequency of different cohesion types in a variety of text styles. According to their results, lexical cohesion is the most dominant category. The method to identify lexical cohesion in text was by mean of lexical chains (see Section 3.6.1 for more details).

[4] The 2-norm (Euclidean Distance) between two vectors of word sentence (gold standard sentence against automatic summary one)

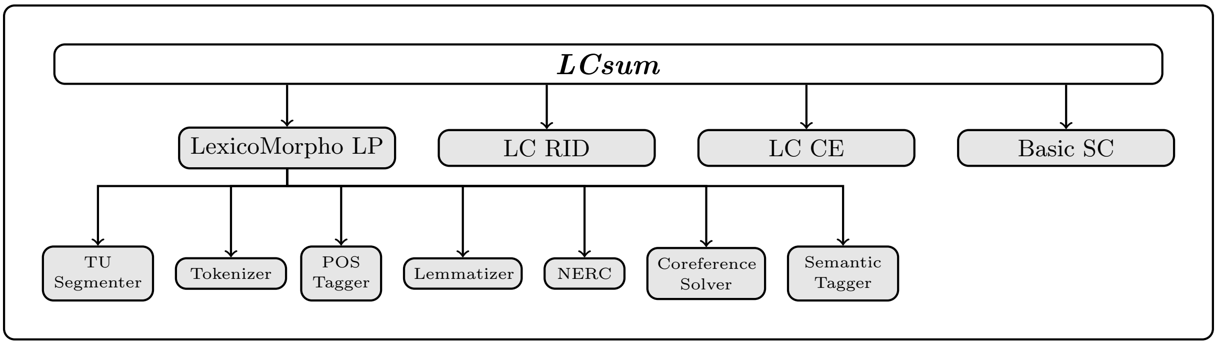


Figure 4.2: *LCsum* FEMsum instantiation process.

As can be seen in Figure 4.2, the system performs four steps, each one based on a simple processor:

1. **LexicoMorpho LP**: The text is linguistically processed, providing the *sent* lexico-morphological representation for further processes. The modules of this component have been described in Section 3.4. The document can be segmented, at different granularity levels (paragraph, sentence, clause or chunks) depending on the application. For the experiments reported in this section, text is segmented at paragraph level and coreference is resolved.

2. **Lexical Chain based RID**: As said in Section 3.6.1 LexChain RID uses the Lexical Chainer to compute lexical chains and identify the strong ones. Some parameters were left unaltered for this first experiment: only *strong* or *extra-strong* kind of WN (or EWN) relations between chain candidates were considered and the text was represented only by *Strong* chains. However, the type of chain to be produced was not the same for all runs. In a first run only common noun chains were considered, in a second run proper name chains were added, and in the third run coreference was solved considering personal pronouns as candidate words, to be related with their references.

3. **Lexical Chain based CE**: In a first experiment, different heuristics were tested. Significant TUs were ranked and selected as described in Section 3.7.1. Two heuristic schemata were tested: *heuristic 1* ranks as most relevant the first TU crossed by a *Strong* chain, while *heuristic 2* ranks highest the TU crossed by the maximum of *Strong* chains.

4. **Basic SC**: Producing summaries of 10% of compression taking into account the relevance score established by the previous component.

Results when comparing several configurations of *LCsum* and the two baselines, *Lead* and *SweSum*, are presented in Table 4.1. The first column in the table shows the main *LCsum*

	Precision	**Recall**	**Cosine**
Lead	.95	.85	.90
SweSum	.90	.81	.87
	Heuristic 2		
Lex. Chains	.71	.72	.79
Lex. Chains + PN Chains	.73	.74	.81
Lex. Chains + PN Chains + coRef Chains	.70	.71	.78
Lex. Chains + PN Chains + coRef Chains + 1st TU	.82	.82	.86
	Heuristic 1		
Lex. Chains	.82	.81	.85
Lex. Chains + PN Chains	.85	.85	.88
Lex. Chains + PN Chains + coRef Chains	.83	.83	.87
Lex. Chains + PN Chains + coRef Chains + 1st TU	.88	.88	.90

Table 4.1: *LCsum* performance when summarizing agency news.

parameters governing each trial: simple lexical chains, lexical chains augmented with proper name and coreference chains, and finally giving special weighting to the 1st TU because of global document structure applicable to the journalistic genre.

As it can be seen in the first row of Table 4.1, the lead achieves the best results, with almost the best possible score. This is due to the pyramidal organization of the journalistic genre that causes most relevant information to be placed at the beginning of the text. Consequently, any heuristic assigning more relevance to the beginning of the text will achieve better results in this kind of genre. This is the case for the default parameters of *SweSum* (second row in Table 4.1) and *heuristic 1* (last row of the table) .

Adapting the system to summarize Catalan documents

As said in Section 4.1.1, to evaluate the system when dealing with Catalan documents, the Spanish test corpus was automatically translated. Table 4.2 shows the Precision, Recall and Cosine obtained when summarizing Catalan documents (first row) or Spanish documents (last row). The *LCsum* instantiated in this experiment to summarize Catalan and Spanish documents computes common noun, proper name and NE to be taken into account as chain members. Considering the results obtained in previous experiments, coreference chains were not taken into account, and *heuristic 1* was applied.

	Precision	**Recall**	**Cosine**
Catalan	0.83	0.83	0.86
Spanish	0.85	0.85	0.88

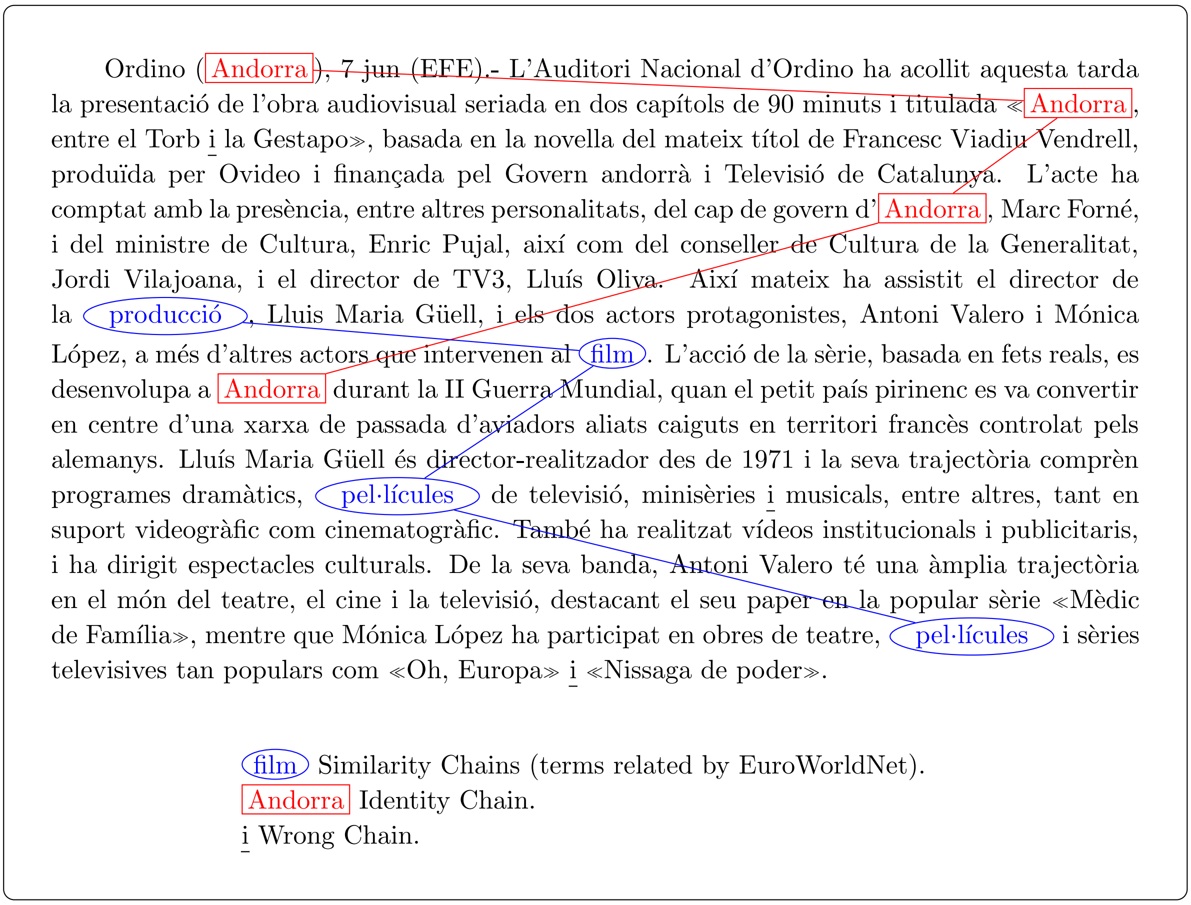
Table 4.2: *LCsum* performance when summarizing the same documents in Catalan or Spanish.

It can be observed that the results obtained for Catalan were somewhat lower than the ones obtained when summarizing Spanish documents. In order to explain this difference, the quality of the translated documents was checked to analyze to what extent the quality of the summary could be affected by the fact of using a MT system. The main drawback found in the Catalan translation was that, in a surprising decision of the MT system, proper names, NEs and acronyms had been translated as common nouns, if a suitable common noun existed. Table 4.3 shows some samples of this phenomenon. However, since the quantitative impact of this error is very limited, despite of the qualitative relevance of these examples, we have considered that it should not affect the quality of the final summary.

	Original	**Translation**
	Francisco Arias Milla	Francisco **Àries** Milla
Proper Names	Cuauhtémoc Cardenas	Cuauhtémoc **Moradenques**
	Carlos Ruiz Sacristan	Carlos Ruiz **Sagristà**
NEs	Trotamundos de Carabobo	**Rodamón** de Carabobo
(baseball team)	Los Padres de San Diego	**Els Pares de St Dídac**
Acronims	(EFE)	(EFA)
	(PAN)	(PA)

Table 4.3: Example of Spanish Proper Names, NEs or acronyms translated as a common nouns in Catalan.

On the other hand, we analyzed each document from which different sentences were selected as a summary for Catalan or for Spanish. Summaries are different mainly when different lexical chains are recognized in Catalan than in Spanish. Table 4.2 reflects the fact that lexical chain quality depends on the propagated linguistic process error. The linguistic tools used for Catalan usually perform a little bit worse than the ones for Spanish. The linguistic processes that most negatively affect the Lexical Chainer component are the POS tagging and the NERC.



Ordino (Andorra), 7 jun (EFE).- L'Auditori Nacional d'Ordino ha acollit aquesta tarda la presentació de l'obra audiovisual seriada en dos capítols de 90 minuts i titulada ≪Andorra, entre el Torb i la Gestapo≫, basada en la novella del mateix títol de Francesc Viadiu Vendrell, produïda per Ovideo i finançada pel Govern andorrà i Televisió de Catalunya. L'acte ha comptat amb la presència, entre altres personalitats, del cap de govern d'Andorra, Marc Forné, i del ministre de Cultura, Enric Pujal, així com del conseller de Cultura de la Generalitat, Jordi Vilajoana, i el director de TV3, Lluís Oliva. Així mateix ha assistit el director de la producció, Lluis Maria Güell, i els dos actors protagonistes, Antoni Valero i Mónica López, a més d'altres actors que intervenen al film. L'acció de la sèrie, basada en fets reals, es desenvolupa a Andorra durant la II Guerra Mundial, quan el petit país pirinenc es va convertir en centre d'una xarxa de passada d'aviadors aliats caiguts en territori francès controlat pels alemanys. Lluís Maria Güell és director-realitzador des de 1971 i la seva trajectòria comprèn programes dramàtics, pel·lícules de televisió, minisèries i musicals, entre altres, tant en suport videogràfic com cinematogràfic. També ha realitzat vídeos institucionals i publicitaris, i ha dirigit espectacles culturals. De la seva banda, Antoni Valero té una àmplia trajectòria en el món del teatre, el cine i la televisió, destacant el seu paper en la popular sèrie ≪Mèdic de Família≫, mentre que Mónica López ha participat en obres de teatre, pel·lícules i sèries televisives tan populars com ≪Oh, Europa≫ i ≪Nissaga de poder≫.

film Similarity Chains (terms related by EuroWorldNet).
Andorra Identity Chain.
i Wrong Chain.

Figure 4.3: Example of an original agency news document in Catalan from EFE with the corresponding lexical chains.

Figure 4.3 is an example of a Catalan EFE agency piece of news. It is a candidate document to be summarized. In this piece of news we can see an example of *similarity* chain ("film" in Figure 4.3) and another of *identity* chain ("Andorra" in the figure). Moreover, the figure shows an example of incorrect chain, with the word "i", which is the conjunction "*and*". In this document all apparitions of “i” should be tagged with the “conjunction” POS tag, however, sometimes it has been tagged as a common noun as if in the document the “i” was referring to the vowel, which was not the case.

Adding coherence textual properties

One of the drawbacks of lexical chains is that they are insensitive to the non-lexical structure of the text, such as their rhetorical, argumentative or document structure. For example, they don’t take into account the position of the elements of a chain within the argumentative line of the discourse. Obtaining and using the discourse structure of a text can contribute to overcome this drawback.

Coherence defines a discourse macro-level structure by relations between text segments (elaboration, cause, explanation..., see an example in Figure 4.4). The rhetorical structure is represented as a hierarchical structure with minimal discourse units (segments) and relations between them (coherence or rhetorical relations). The assumption is that the most important segments are placed at the top of the hierarchical structure. In the example, the top segment is “**Mars experiences frigid weather conditions**”. While the segment situated at the bottom, “**50% farther from the sun than Earth**”, is the less important one.

As follows, cohesion and coherence account for complementary aspects of the discourse structure of a text. On one hand, lexical chains account for the linear distribution of content in a text, considering as most relevant fragments of text those where most of the identified content lines are represented. On the other hand, discourse structure relations provide a hierarchical structure of the same text.

Some theoretical approaches to discourse processing, such as (Polanyi 1988), give an integrated account of a number of linguistic levels, including these two. However, cohesion-based discourse models and coherence-based ones have usually worked separately for general-purpose NLP applications.

Details about the lexical chain text representation are given in Section 3.6.1 and the rhetorical representation used in the reported experiments is the same as the one detailed in the Rhetorical Argumentative Compressor (see Section 3.8.1). Figure 4.5 presents a Spanish document from the HERMES corpus represented with lexical chains, while in Figure 4.6 the rethorical information of the same document is reflected.

In Figure 4.5 two strong chains have been detected: the *similarity* chain “empresarios” (“*businessmen*”) and the *identity* chain “Generalitat” (Catalan government). Moreover, the most relevant fragment is the one where most of the identified content is represented (see the TU in the box in Figure 4.5). However, looking at Figure 4.6, it can be seen that the most relevant segment detected by the lexical chains is in grey, which indicates less important rhetorical segments, while

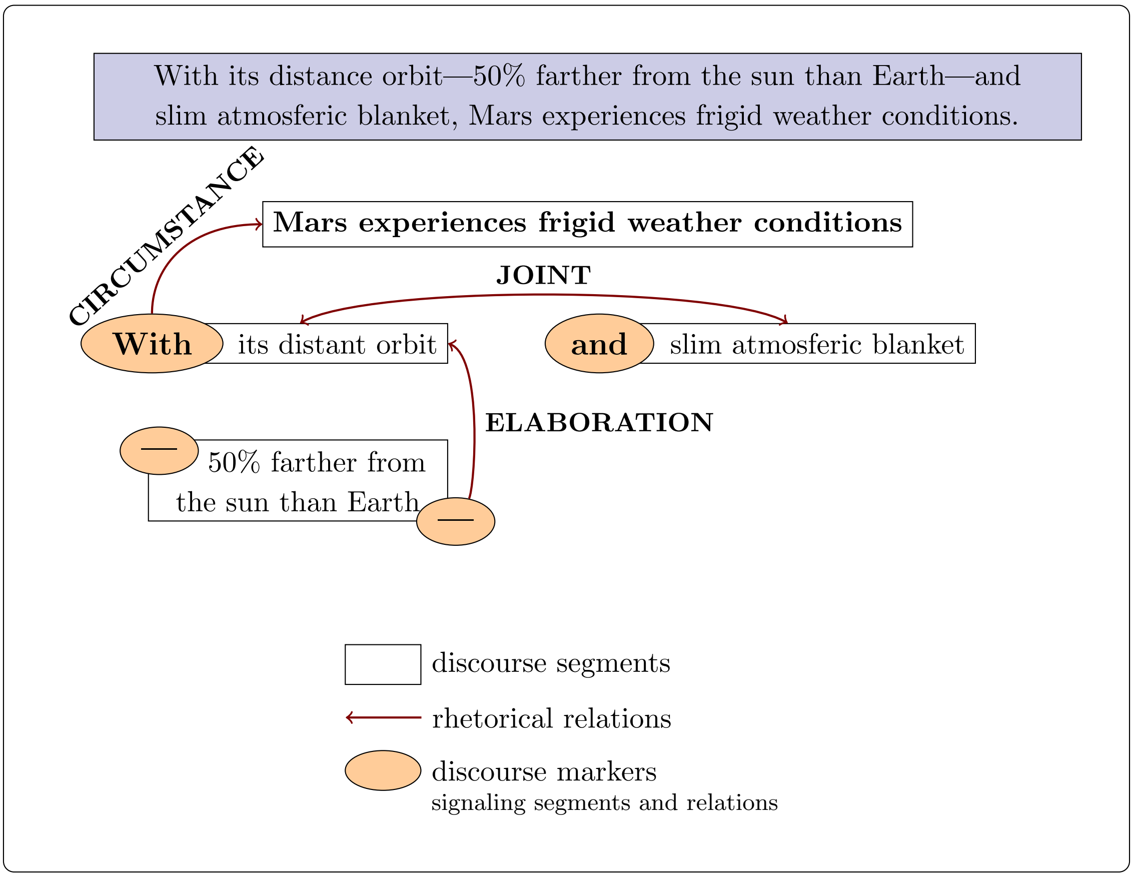

With its distance orbit—50% farther from the sun than Earth—and slim atmosferic blanket, Mars experiences frigid weather conditions.
CIRCUMSTANCE
Mars experiences frigid weather conditions
JOINT
With
its distant orbit
and
slim atmosferic blanket
ELABORATION
—
50% farther from the sun than Earth
—
discourse segments
rhetorical relations
discourse markers
signaling segments and relations

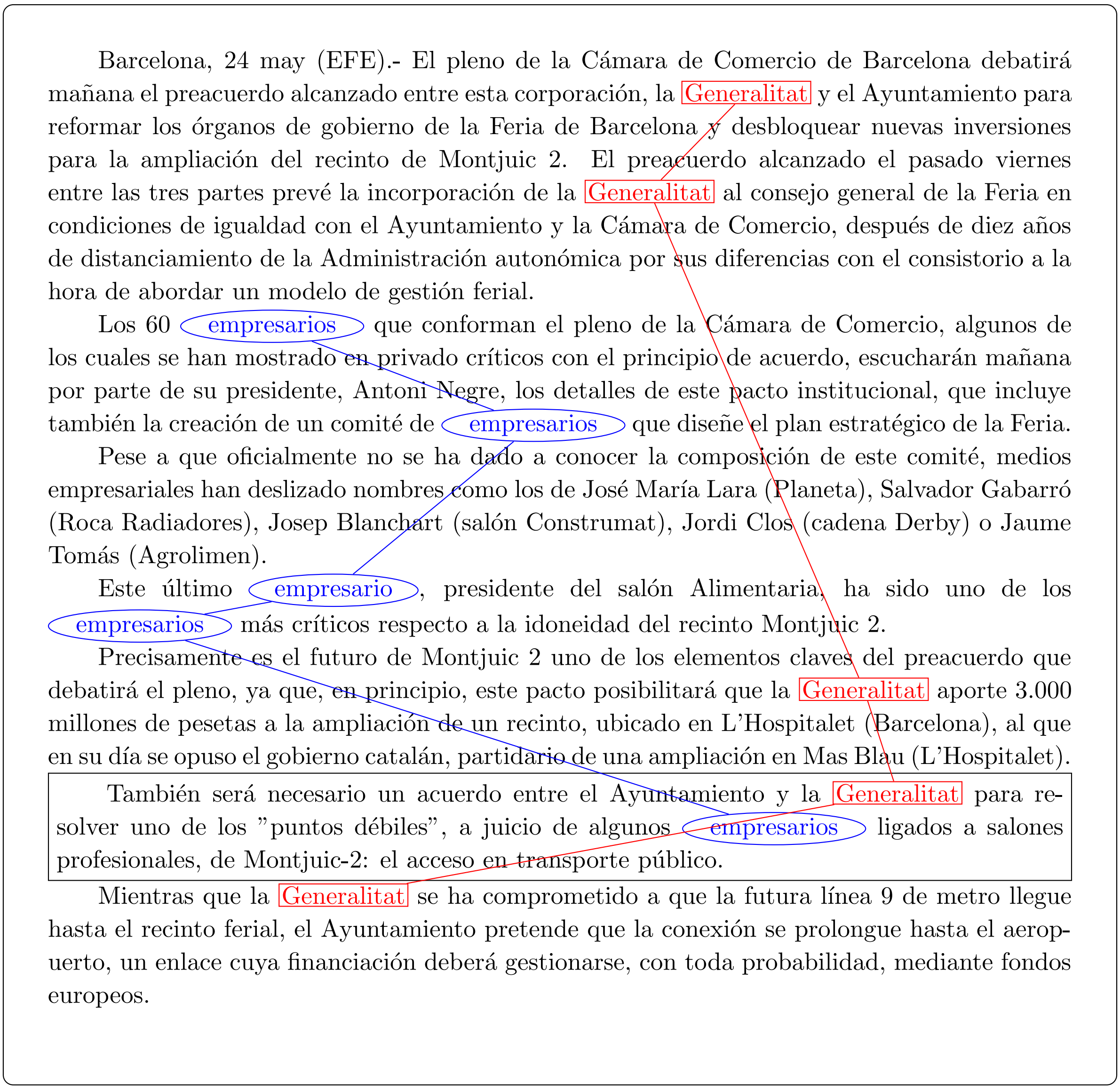
Figure 4.4: Example of the rhetorical structure of a sentence.

text containing central segments is underlined, and discourse markers are circled.

In order to improve the performance of the *LCsum* approach, several experiments have been carried out to study how the cohesion-based *LCsum* can successfully be enriched with discourse aspects, *LCsum(DiscStr)*.

First of all, in (Alonso and Fuentes 2002) we present an initial collaboration between two complementary approaches, *Lexical Chains* and *Rhetorical Structure*. For this first experiment, the two approaches worked sequentially. Rhetorical structure allows us to detect in advance non-relevant parts of text and to remove them before applying *LCsum*.

In a second experiment (Alonso and Fuentes 2003), *LCsum(DiscStr)* takes into account the rhetorical and argumentative structure obtained via Discourse Marker (DM). As can be seen in Figure 4.7 (right), the previously described *LCsum* approach is enhanced with discourse structural information, in order to test whether taking into account the structural status of the



Barcelona, 24 may (EFE).- El pleno de la Cámara de Comercio de Barcelona debatirá mañana el preacuerdo alcanzado entre esta corporación, la Generalitat y el Ayuntamiento para reformar los órganos de gobierno de la Feria de Barcelona y desbloquear nuevas inversiones para la ampliación del recinto de Montjuic 2. El preacuerdo alcanzado el pasado viernes entre las tres partes prevé la incorporación de la Generalitat al consejo general de la Feria en condiciones de igualdad con el Ayuntamiento y la Cámara de Comercio, después de diez años de distanciamiento de la Administración autonómica por sus diferencias con el consistorio a la hora de abordar un modelo de gestión ferial.

Los 60 empresarios que conforman el pleno de la Cámara de Comercio, algunos de los cuales se han mostrado en privado críticos con el principio de acuerdo, escucharán mañana por parte de su presidente, Antoni Negre, los detalles de este pacto institucional, que incluye también la creación de un comité de empresarios que diseñe el plan estratégico de la Feria.

Pese a que oficialmente no se ha dado a conocer la composición de este comité, medios empresariales han deslizado nombres como los de José María Lara (Planeta), Salvador Gabarró (Roca Radiadores), Josep Blanchart (salón Construmat), Jordi Clos (cadena Derby) o Jaume Tomás (Agrolimen).

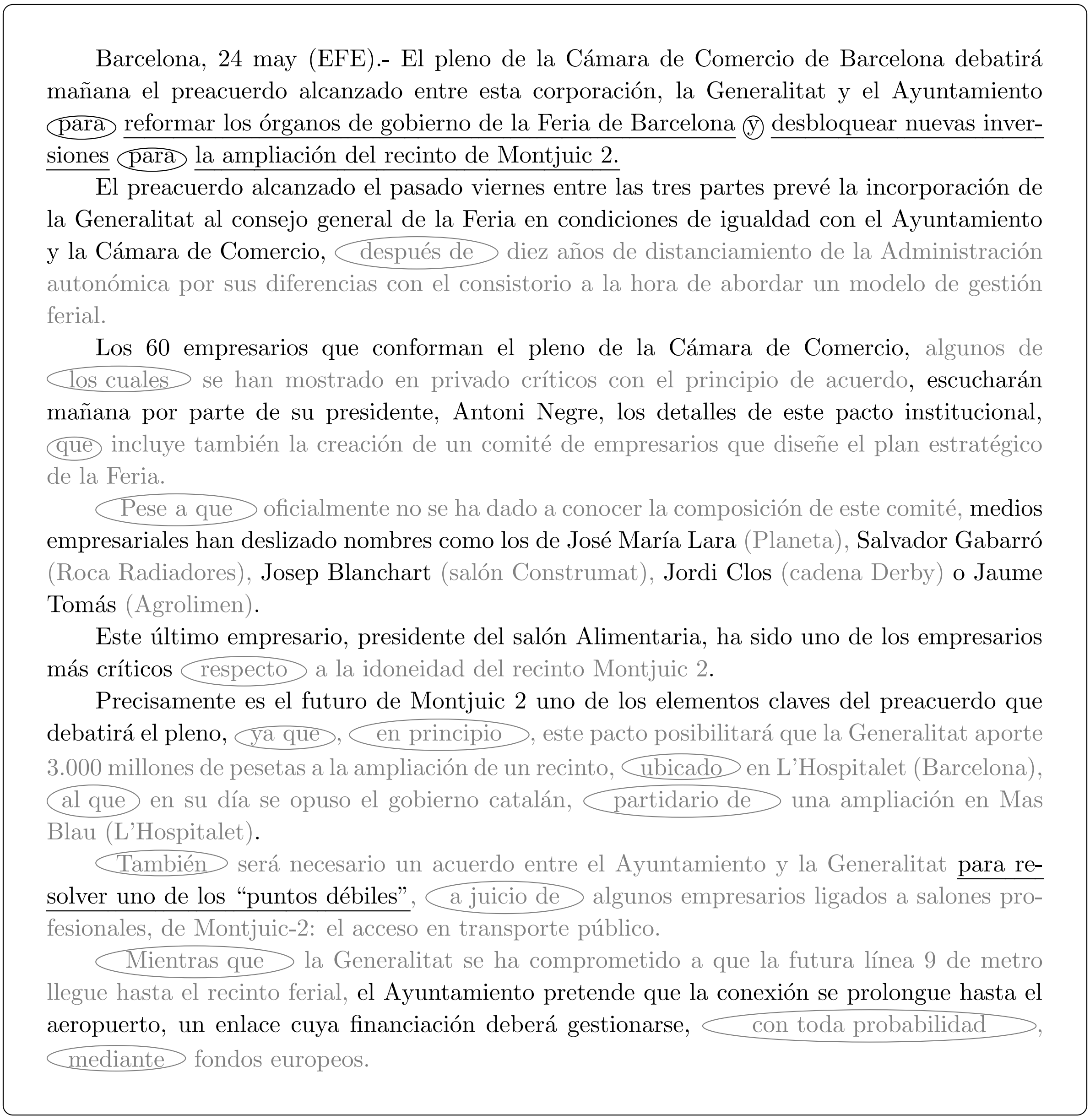
Este último empresario, presidente del salón Alimentaria, ha sido uno de los empresarios más críticos respecto a la idoneidad del recinto Montjuic 2.

Precisamente es el futuro de Montjuic 2 uno de los elementos claves del preacuerdo que debatirá el pleno, ya que, en principio, este pacto posibilitará que la Generalitat aporte 3.000 millones de pesetas a la ampliación de un recinto, ubicado en L'Hospitalet (Barcelona), al que en su día se opuso el gobierno catalán, partidario de una ampliación en Mas Blau (L'Hospitalet).

También será necesario un acuerdo entre el Ayuntamiento y la Generalitat para resolver uno de los "puntos débiles", a juicio de algunos empresarios ligados a salones profesionales, de Montjuic-2: el acceso en transporte público.

Mientras que la Generalitat se ha comprometido a que la futura línea 9 de metro llegue hasta el recinto ferial, el Ayuntamiento pretende que la conexión se prolongue hasta el aeropuerto, un enlace cuya financiación deberá gestionarse, con toda probabilidad, mediante fondos europeos.

Figure 4.5: A text represented by lexical chains. Boxes are used for the identity chain members and circles for similarity chain members.



Barcelona, 24 may (EFE).- El pleno de la Cámara de Comercio de Barcelona debatirá mañana el preacuerdo alcanzado entre esta corporación, la Generalitat y el Ayuntamiento para reformar los órganos de gobierno de la Feria de Barcelona y desbloquear nuevas inversiones para la ampliación del recinto de Montjuic 2.

El preacuerdo alcanzado el pasado viernes entre las tres partes prevé la incorporación de la Generalitat al consejo general de la Feria en condiciones de igualdad con el Ayuntamiento y la Cámara de Comercio, después de diez años de distanciamiento de la Administración autonómica por sus diferencias con el consistorio a la hora de abordar un modelo de gestión ferial.

Los 60 empresarios que conforman el pleno de la Cámara de Comercio, algunos de los cuales se han mostrado en privado críticos con el principio de acuerdo, escucharán mañana por parte de su presidente, Antoni Negre, los detalles de este pacto institucional, que incluye también la creación de un comité de empresarios que diseñe el plan estratégico de la Feria.

Pese a que oficialmente no se ha dado a conocer la composición de este comité, medios empresariales han deslizado nombres como los de José María Lara (Planeta), Salvador Gabarró (Roca Radiadores), Josep Blanchart (salón Construmat), Jordi Clos (cadena Derby) o Jaume Tomás (Agrolimen).

Este último empresario, presidente del salón Alimentaria, ha sido uno de los empresarios más críticos respecto a la idoneidad del recinto Montjuic 2.

Precisamente es el futuro de Montjuic 2 uno de los elementos claves del preacuerdo que debatirá el pleno, ya que, en principio, este pacto posibilitará que la Generalitat aporte 3.000 millones de pesetas a la ampliación de un recinto, ubicado en L'Hospitalet (Barcelona), al que en su día se opuso el gobierno catalán, partidario de una ampliación en Mas Blau (L'Hospitalet).

También será necesario un acuerdo entre el Ayuntamiento y la Generalitat para resolver uno de los "puntos débiles", a juicio de algunos empresarios ligados a salones profesionales, de Montjuic-2: el acceso en transporte público.

Mientras que la Generalitat se ha comprometido a que la futura línea 9 de metro llegue hasta el recinto ferial, el Ayuntamiento pretende que la conexión se prolongue hasta el aeropuerto, un enlace cuya financiación deberá gestionarse, con toda probabilidad, mediante fondos europeos.

Figure 4.6: A text represented by rhetorical structure. Grey text contains rhetorically subordinate text segments, while underlined text contains central segments and discourse markers are in circles.

TU where a chain member occurs can improve the ranking of lexical chains.

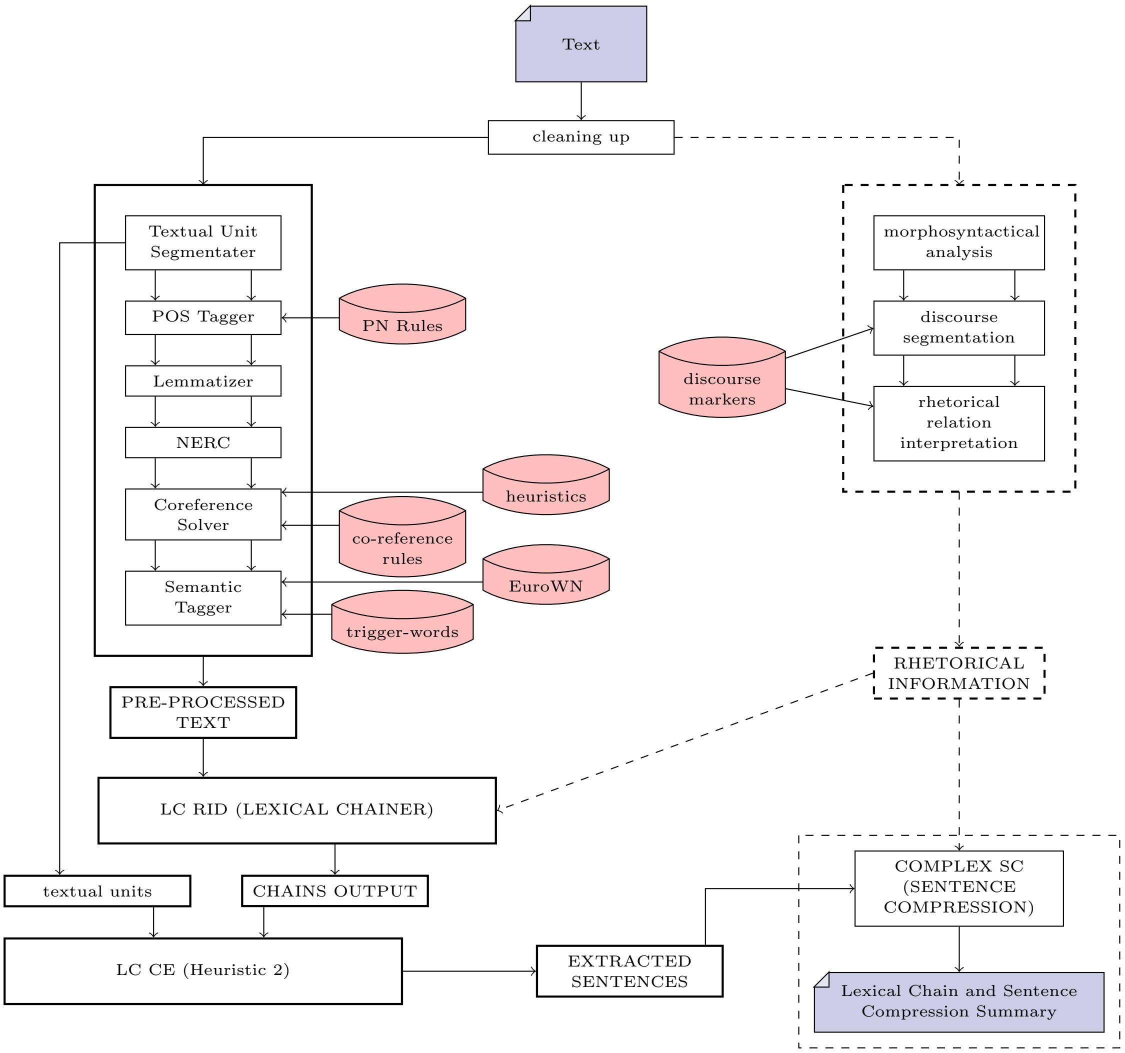


Figure 4.7: *LCsum (DiscStr)*, integration of discursive information: lexical chains (left) and discourse structural (right).

LCsum (DiscStr) works in four steps:

1. **LexicoMorpho LP**: The *sent* lexico-morphological representation is produced and the document is segmented at discourse level or at sentence level. In the first case, the text has previously been enriched with the rhetorical representation, as detailed in the Rhetorical Argumentative Compressor (see Section 3.8.1).

2. **Lexical Chain based RID**: In addition to sentences, discourse segments are allowed to be the LEXICAL CHAINER input TUs, thus allowing a finer granularity level than sentences. *Strong* or *extra-strong* kind of relations between chain candidates are considered. Candidate chain members are common nouns, proper names, and NEs.

3. **Lexical Chain based CE**: *Heuristic 2*, ranking highest the TU crossed by the maximum of *Strong* chains, is set.

4. **Complex SC**: In the reported experiments two kinds of TU are considered as input of the SC component. If the input is a set of discourse segments, the Basic SC is used to produce extract based summaries of about 10% of the original document length. However, when the input TUs are sentences, the instantiated SC includes the Rhetorical Argumentative Compressor detailed in Section 3.8.1. In the compression process, lexical chain members are taken into account. Some rhetorically subordinate text segments from the most relevant candidate TUs are removed and the final summary is composed by adding only central segments, until achieving a summary of the 10% with respect the original document.

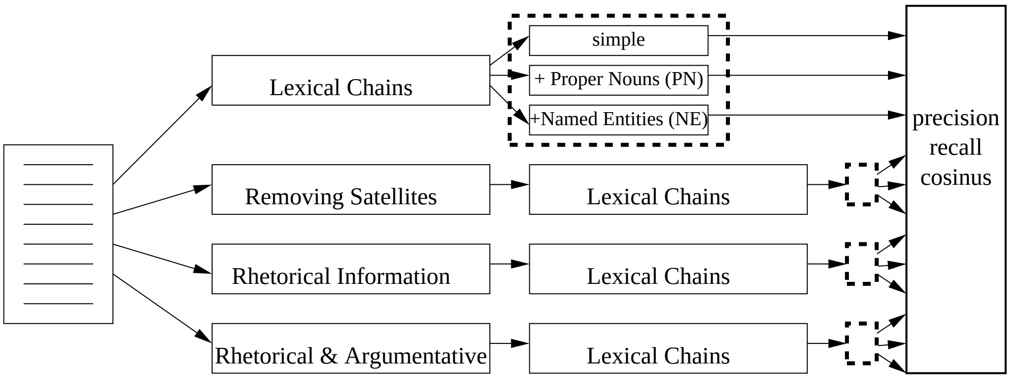


Figure 4.8: Experiments to assess the impact of discourse structure on lexical chain members.

A number of experiments were carried out, in order to assess the impact of considering discursive structural context of lexical chain members (see Figure 4.8).

The results of integrating lexical chains with discourse structural information can be seen in Table 4.4. Following the design sketched in Figure 4.8, the *LCsum* performance was first evaluated on a text where satellites had been removed. The rhetorical analysis identifies nucleus-satellite relations, satellites are less relevant because of their rhetorical subordination to nuclei.

	Precision	**Recall**	**Cosine**
Sentence Compression + Lexical Chains			
Sentence Compression + Lexical Chains + PN Chains	.74	.75	.70
Sentence Compression + Lexical Chains + PN Chains + 1st TU	.86	.85	.76
Rhetorical Information + Lexical Chains			
Rhetorical Information + Lex. Chains + PN Chains	.74	.76	.82
Rhetorical Information + Lex. Chains + PN Chains + 1st TU	.83	.84	.86
Rhetorical + Argumentative + Lexical Chains			
Rhetorical Information + Argumentative + Lex. Chains + PN Chains	.79	.80	.84
Rhetorical Information + Argumentative + Lex. Chains + PN Chains + 1st TU	.84	.85	.87

Table 4.4: Results of the integration of lexical chains and discoursive structural information

As stated by (Brunn et al. 2001b) and (Alonso and Fuentes 2002), removing satellites slightly improves the relevance assessment of the *LCsum*.

Secondly, distinctions between rhetorical information and argumentative information have been considered, since the first identifies mainly unimportant parts of text and the second identifies both important and unimportant. As can be seen in Table 4.4, identifying satellites instead of removing them yields only a slight improvement on recall (from .75 to .76), but significantly improves cosine (from .70 to .82).

When argumentative information is provided, an improvement of 0.05 in performance is observed in all three metrics in comparison to removing satellites. As can be expected, ranking the first TU higher results in better measures, because of the nature of the genre.

Finally, intra-sentential satellites of the best summary obtained by lexical chains were removed, increasing compression of the resulting summaries from an average 18.84% for lexical chain-based summaries to a 14.43% for summaries which were sentence-compressed. However, these summaries have not been evaluated with the MEADeval package because no gold standard was available for TUs smaller than paragraphs.

4.2 News Stories Headline Extraction

The *MLsum* FEMsum was first instantiated to participate in the DUC 2003 SDS task. The aim of this task was to produce headlines, very short summaries, 10-word single document summaries for pieces news of written English.

The *MLsum*, described in Section 4.2.2, is aimed to generalize the *LCsum* analyzed in Section 4.1.2 by taking into account relevant information other than lexical cohesion. *MLsum* applies Machine Learning (ML) techniques to produce summaries by TU extraction. An optional module, the Full Parsing Sentence Compressor, was developed with the aim of producing grammatical headlines. *MLsum(HL)*, the approach evaluated in DUC 2003 (Fuentes et al. 2003) uses this module in the SC component.

To be able to evaluate evolving prototypes of the summarizer, Section 4.2.1 presents a method for re-using the human judgments on summary quality provided by the DUC 2002 contest. The score to be assigned to automatic summaries is calculated as a function of the scores manually assigned to the most similar summaries for the same document. This approach enhances the standard n-gram based evaluation of automatic summarization systems by establishing similarities between *extractive* (vs. *abstractive*) summaries and by taking advantage of the big quantity of evaluated summaries available from the DUC contest. The utility of this method (Alonso et al. 2004) is exemplified by the evaluation of the improvements achieved on subsequent *MLsum(HL)* versions (see details in Section 4.2.2).

4.2.1 Re-using DUC data for evaluation

In order to avoid the dependency on human judgments for evaluating improvements in our system or other summarization approaches that were introduced after submission to the DUC contest, this section describes an automatic evaluation method. We propose to automatically evaluate a new system by using as knowledge sources both manual and automatic DUC summaries as well as the scores manually assigned to them.

Our methodology goes beyond ROUGE, presented in Section 2.3, taking advantage of two facts:

- n-gram overlap is more adequate to account for similarities between *extractive* summaries than between *abstractive* summaries
- a big quantity of human-made judgements on summary quality is available for a big number of extractive summaries, produced by NIST assessors for DUC

ROUGE establishes comparisons between automatic, mostly *extractive* summaries, and human, *abstractive* summaries. It can be expected that similarities between pairs of *extractive* summaries are even better represented by n-gram overlap, because the variability in linguistic realization is lower in extractive summaries than in human-generated summaries, since words used to produce the summaries all come from the same original text.

As for the big quantity of judgements, all summaries submitted by every system to DUC are available for every participant, together with the score assigned to them. It can be expected, then, that word-based measures do account for similarity between automatic summaries that have received comparable scores in DUC, because most of the participating systems took an extraction-based approach.

We propose to approximate human scoring of a new summary by weighting the scores assigned by DUC assessors to similar summaries submitted to DUC contest. More precisely, our scoring simply computes the weigthted average of the scores assigned by human judges of the N most similar summaries to the summary to be evaluated. Similarity between the new summary and the evaluated summaries is calculated by unigram overlap.

For testing our proposal we applied this methodology (setting N to 3) to the systems participating in Task 1 of DUC 2003. We used the scores (ranging from 0..1) manually assigned by NIST assessors for:

- *DUC Coverage*: how much of a model summary's content was expressed by a system-generated peer summary;
- *DUC Length-Adjusted Coverage (LAC)*: the ability of a system to produce a summary

shorter than the predefined target length devise a combined measure of coverage and compression.

We calculated the approximated scores as follows:

$$score = \frac{\sum_{i=1}^{3} v_i s_i^2}{\sum_{i=1}^{3} s_i^2} \tag{4.1}$$

where
s is the similarity between a summary and the summary to be scored, by unigram overlap normalized by the size of the strings to be compared. v is the score for *Coverage* or *LAC* assigned to that summary.

The scores assigned automatically to a given summary correlate very well with the scores assigned to the three other most similar summaries by unigram overlap: the correlation coefficient amounts to 0.99 between approximated and manual scores both for *Coverage* and *LAC*.

4.2.2 A mixed approach to Headline Extraction for DUC task

This section presents the approach to Headline Extraction presented at DUC 2003. It consists of a sentence extractor ML-based approach, *MLsum*, plus an optional Simplification module that produces grammatical headlines, *MLsum(HL)*.

The approaches presented in this section carry out the summarization process as described in Figure 4.9. The FEMsum architecture components are instantiated as follows:

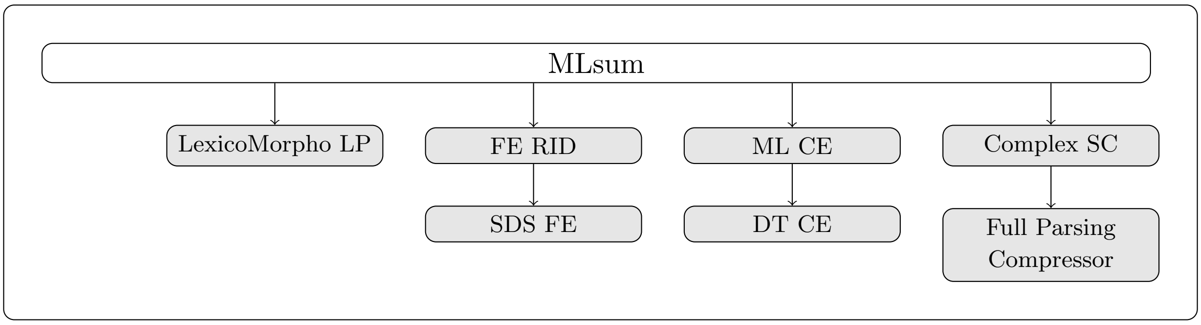


Figure 4.9: *MLsum* FEMsum instantiation process.

1. **LexicoMorpho LP**: The document is processed to obtain the *sint* syntanctic representation of each TU. In the reported experiments, TUs are sentences.

2. **FeatExt RID**: TUs are enriched with features relevant to the task implementing the SDS Feature Extr with the DT Feature Extr as described in Section 3.6.2. This component uses the Lexical Chainer to compute one of the features, which indicates the number of strong chains in each TU.

3. **ML CE**: each TU is classified by the DT CE described in Section 3.7.2 as belonging to the summary or not, according to its features and a set of classification rules automatically learned from a training corpus. A confidence score is assigned to each decision, based on the confidence associated to the rule applied, and the set of possible summary TUs is ranked accordingly. For SDS, the classification task is carried out by means of decision rules automatically extracted from a C4.5 Decision Tree (DT) algorithm.

4. **Complex SC**: As detailed in Section 3.8.1, in *MLsum(HL)*, once the set of possible summary TUs is determined, the Full Parsing Sentence Compression module applies several rules to obtain a grammatical headline. The ranking of TUs can be parametrized so that a TU can be assigned a different relative scoring if it is crossed by a strong chain, by a NE chain or by a coreference chain.

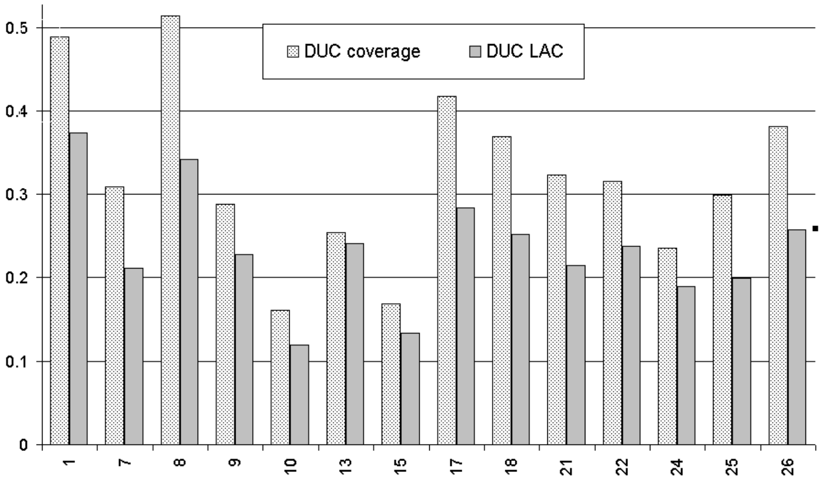


Figure 4.10: Manual evaluations for *Coverage* and *Length-Adjusted Coverage* for the DUC 2003 participant systems [7 - 26] and the DUC Baseline [1].

Figure 4.10 displays results for *Coverage* and *Length-Adjusted Coverage (LAC)* of automatic systems participating in DUC 2003 Task1. The results of the *MLsum(HL)*, number 24, were somewhat dissapointing. In order to guide future improvements on the *MLsum(HL)*, the obtained results and the quality judgements provided by NIST were analyzed. The system identified with number 8 was discarded because the summaries were much longer than the required 10 words.

To begin with, it must be said that a summary was provided for only 91% of the 624 documents to be summarized, due to restrictive heuristics for choosing TUs to be included in the summary, and lack of robustness of compression rules.

With respect to length, our mean is quite in target: 8.33 words. However, while few of our headlines were longer than the mean, a considerably bigger number are much shorter than 10 words. Since grammaticality was priorized, these very short summaries contained many more grammatical words than content words.

To identify weak points, an error analysis of the two main components was carried out separately. First, the performance of the ML CE component was analyzed with respect to *Coverage* with the methodology presented in Section 4.2.1. Then, the effects on coverage of the Sentence Compression process were also analyzed.

Coverage analysis of the Content Extractor

With the aim of analyzing if the sentence selected by the ML CE component was adequate or not, we applied the procedure described in Section 4.2.1 to the complete text of those sentences compressed to produce the headline.

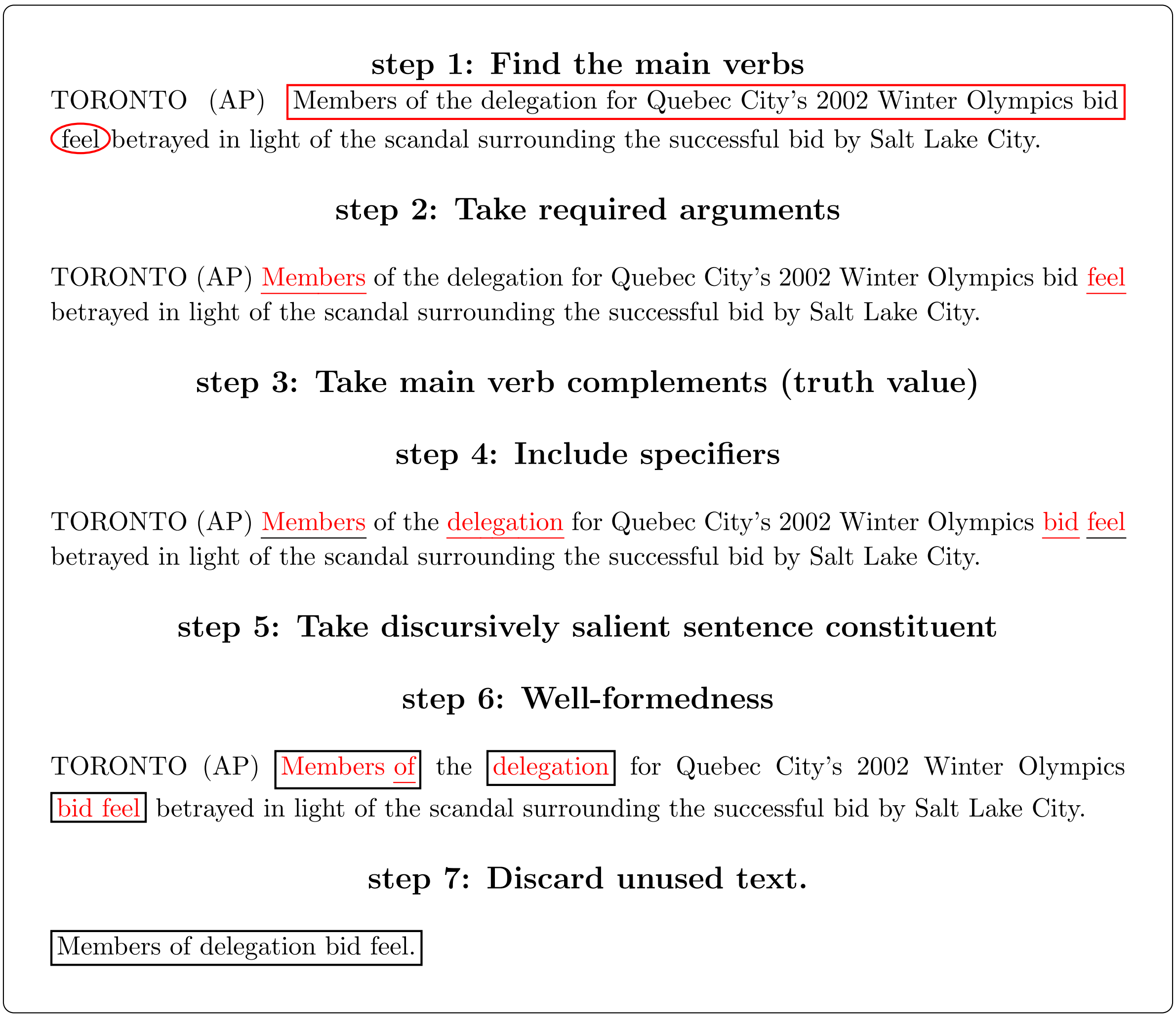
From the 564 submitted summaries, 211 were manually assigned 0 for coverage at DUC. From these headlines, 210 were automatically assigned less than 0.1 approximate coverage. However, only 97 of the original sentences (before compression) from which coverage 0 summaries were manually assigned obtained less than 0.1 of approximate coverage. This means than in more than 50% of the cases, the compression process caused a total loss of summary content coverage respect the content coverage of the original sentence. Possible causes for this decrease in coverage are analyzed in next section.

Effects of Sentence Compression

To analyze the reduction in coverage caused by the Sentence Compression module used in our DUC 2003 participation, a set of 84 sentences was studied, 58 of which were considered unsatisfactory, either from the point of view of coverage or grammaticality, although the grammatical aspect was not reflected in DUC results.

In more than half of the cases (34 sentences), the loss of coverage in reducing the original sentence was due to an inadequate treatment of highly informative elements, like NEs or words which are member of a lexical chain. This is a shortcoming of compression rules, which were based exclusively in structural information, and did not take into account the lexical status of the elements in sentence constituents. Additionally, compression rules caused grammaticality errors in 26 sentences, and parsing errors affected 15 sentences.

Figure 4.11 presents the steps of the Full Parsing Sentence Compression SC component presented in Section 3.8.1. This component is applied to a sentence previously extracted as relevant by the ML CE.



step 1: Find the main verbs

TORONTO (AP) Members of the delegation for Quebec City's 2002 Winter Olympics bid feel betrayed in light of the scandal surrounding the successful bid by Salt Lake City.

step 2: Take required arguments

TORONTO (AP) Members of the delegation for Quebec City's 2002 Winter Olympics bid feel betrayed in light of the scandal surrounding the successful bid by Salt Lake City.

step 3: Take main verb complements (truth value)

step 4: Include specifiers

TORONTO (AP) Members of the delegation for Quebec City's 2002 Winter Olympics bid feel betrayed in light of the scandal surrounding the successful bid by Salt Lake City.

step 5: Take discursively salient sentence constituent

step 6: Well-formedness

TORONTO (AP) Members of the delegation for Quebec City's 2002 Winter Olympics bid feel betrayed in light of the scandal surrounding the successful bid by Salt Lake City.

step 7: Discard unused text.

Members of delegation bid feel.

Figure 4.11: Steps of the Full Parser Compressor process.

The resulting compression sentence (see last step of Figure 4.11) does not convey much of the relevant information of the original sentence. The main errors are:

- Bad account of verb argumental structure: since there is no information on the arguments required by the verb *feel*, the adjectival phrase *betrayed* is considered optional by compression rules.
- Parsing errors: the attachment ambiguity of *surrounding*, which depends from *scandal*, is not well resolved, thus it is considered directly depending from the verb and it is assigned the status of an optional verbal adjunct by compression rules.
- Insensitiveness to the lexical status of words: the fact that NEs, like *Quebec City*, *Winter Olympics* or *Salt Lake City*, bear much of the content of the sentence is not captured, because they are not in syntactically salient position (they are not phrasal heads).
- Inadequate treatment of MW expressions: the lack of relevance of the constituent introduced by *in light of* cannot be found because this expression is not considered as a single DM and cannot be treated consequently.

Improving the Sentence Compressor

An improved version of *MLsum(HL)* provides solutions for some of the errors in the results submitted to DUC. Heuristics for choosing the TU to be simplified have been refined. First, units with no content other than authorship, location of issue, etc., are identified and discarded by means of pattern matching. This allows progressively decreasing the minimal required length when the combination of heuristics results too restrictive. Moreover, in case no TU is chosen from the set provided by the classification module, a second set is built with all the units in the document, ranked by order of occurrence.

The DUC-submitted version of the *MLsum(HL)* determined the inclusion in the summary of a sentence constituent relying exclusively on syntactical requirements or discursive particles. In the improved version, each lexical item in the chosen TU has been assigned an informativity status, so that words belonging to a strong lexical chain have been considered most informative, and frequent, nonempty words have been assigned a secondary relevance status. Decisions as to the inclusion of sentence constituents in the summary are now taken considering syntactic, rhetoric and lexical information.

Every one of the presented modifications has been evaluated with the methodology presented in Section 4.2.1. The last objective was to assess whether they introduced significant improvements in the performance of the development system, and include them in the stable version of the system. What we observed was that emphasizing the informativity of summaries always yielded improvements in performance. This seems to be a side-effect of the fact that comparisons are established by unigram overlap of the summaries to be compared. This also explains why producing a list of words as a Sentence Compression output outperforms the scores obtained by the system when using a grammatical Sentence Compression module.

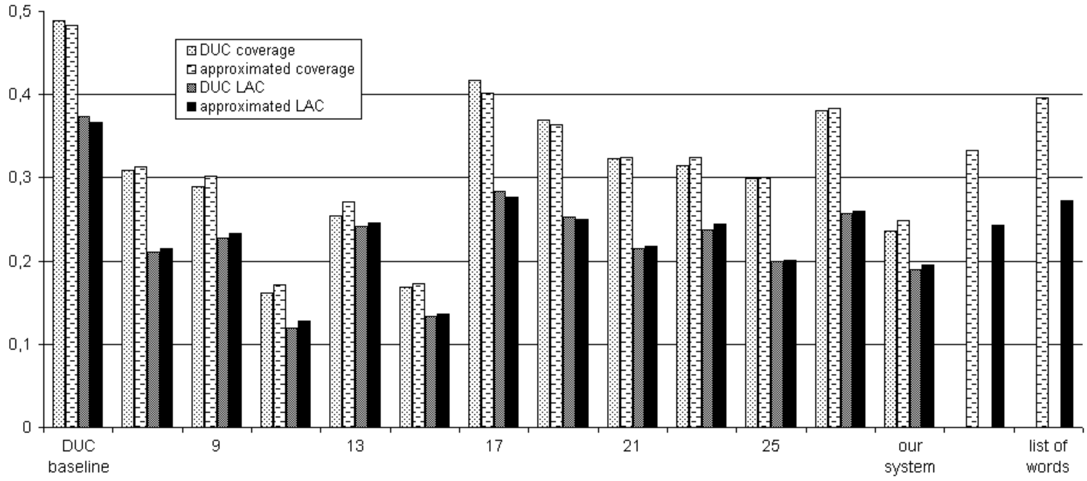


Figure 4.12: For each DUC 2003 participant system 4 values are depicted: the DUC *Coverage*, the approximated *Coverage*, the DUC *LAC* and the approximated *LAC*. For each new evaluated system only approximated values are depicted.

Figure 4.12 displays the results for the DUC manual evaluation and the automatic proposed evaluation. Two kinds of scores are displayed: those assigned manually by NIST assessors for *Coverage* and *Length-Adjusted Coverage (LAC)*, and those approximating the manual scores by the automatic methodology proposed before. From left to right, Figure 4.12 presents the results obtained by: the DUC baseline, the DUC participating systems (*7, 9, 10, 13, 15, 17, 18, 21, 22, 25, 26*), *Our system*, the first prototype of *MLsum(HL)* manually evaluated, the *improved MLsum(HL)* approach and finally, a *list of words* approach. The *list of words* is created by concatenation of the ten most relevant words in the document (strong lexical chain members and frequent words, leaving stopwords out).

When evaluated with the approximated measures and compared with the rest of the DUC 2003 participants, the *list of words* approach achieves competitive results in both approximated measures. But it is not yet as good as the DUC baseline, the original headlines of the documents.

4.3 Conclusions

This chapter details the experiments carried out to set up two FEMsum instantiations to deal with multilingual SDS. *LCsum* uses a cohesion-based representation of the text to detect relevant information. *MLsum* uses a DT based algorithm for scoring TUs according to their likelihood to be included in the final summary.

A corpus to test systems when summarizing Spanish agency news was created in the framework of HERMES project. This corpus has been automatically translated to evaluate the performance when summarizing Catalan documents. It has been observed that the two main shortcomings of this corpus are its small size and the fact that it belongs to the journalistic genre. That means that documents follow a pyramidal organization, with the first sentence as a summary of rest of the document. However, to our knowledge, there is no other corpus for Spanish or Catalan summary evaluation.

Both FEMsum instantiations, *LCsum* and *MLsum*, present portability to a variety of languages, provided there is at least a morphological analyzer and a version of WN (or EWN) available for the language. The performance of the first one has been analyzed for Spanish and Catalan, while the second one participated in the DUC international contest summarizing English news.

Since the properties exploited by *LCsum* are text-bound, they can be considered to have language-wide validity. This means that *LCsum* is domain-independent, though it can be easily tuned to different genres. Moreover, the same approach can be used to summarize a set of documents, written in the same language, MDS, or in different languages, cross-lingual MDS.

LCsum(DiscStr) is a collaborative integration of heterogeneous discursive information that yields an improvement on the representation of source text. However, this enriched representation does not outperform a baseline consisting of taking the first paragraph of the text.

MLsum(HL) uses a linguistically guided compression procedure to produce grammatical headlines. Compression rules are applied to the TU considered more likely to be included in the summary. The first version of this FEMsum instantiation was evaluated in the headline production task of the DUC 2003 obtaining not very good results. In order to improve the performance, a careful analysis of the results was carried out identifying causes of misperformance.

New versions of the compressed rules used by *MLsum(HL)* were evaluated with a method that re-uses the human judgements on summary quality provided by the DUC contest. The score to be awarded to automatic summaries is calculated as a function of the scores assigned manually to the most similar summaries for the same document. An approximate measure of both *coverage* and *LAC* DUC measures are calculated.

We observed that emphasizing the informativity of summaries always yielded improvements in performance. A *list of words* approach achieves competitive results in both approximated measures. But it is not yet as good as the DUC baseline, the original headlines of the documents.