Integrating cohesion and coherence for Automatic Summarization

Laura Alonso i Alemany
GRIAL
Departament de Lingüística General
Universitat de Barcelona
lalonso@lingua.fil.ub.es

Maria Fuentes Fort
Departament d'Informàtica i Matemàtica Aplicada
Universitat de Girona
maria.fuentes@udg.es

Abstract

This paper presents the integration of *cohesive* properties of text with *coherence* relations, to obtain an adequate representation of text for automatic summarization. A summarizer based on Lexical Chains is enchanced with rhetorical and argumentative structure obtained via Discourse Markers.

When evaluated with newspaper corpus, this integration yields only slight improvement in the resulting summaries and cannot beat a dummy baseline consisting of the first sentence in the document. Nevertheless, we argue that this approach relies on basic linguistic mechanisms and is therefore genre-independent.

1 Motivation

Text Summarization (TS) can be decomposed into three phases: analysing the input text to obtain text representation, transforming it into a summary representation, and synthesizing an appropriate output form to generate the summary text.

Much of the early work in summarization has been concerned with detecting relevant elements of text and presenting them in the "shortest possible form". More recently, an increasing attention has been devoted to the adequacy of the resulting texts to a human user. Well-formedness, cohesion and coherence are currently under inspection, not only because they improve the quality of a summary as a text, but also because they can reduce the final summary by reducing the reading time and cost that is needed to process it.

TS systems that performed best in last DUC contest (DUC, 2002) apply template-driven summarization, by information-extraction procedures in the line of (Schank and Abelson, 1977). This approach yields very good results in assessing relevance and keeping well-formedness, but it is dependent on a clearly defined representation of the information need to be fulfilled and, in most cases, also on some regularities of the kind of texts to be summarized.

In more generic TS, genre-dependent regularities are not always found, and template-driven analysis cannot capture the variety of texts. In addition, the information need is usually very fuzzy. In these circumstances, the most reliable source of information on relevance and coherence properties of a text is the source text itself. An adequate representation of that text should account not only for relevant elements, but also for the relations holding between them, in the diverse textual levels. Exploiting the discursive properties of text seems to accomplish both these requirements, since they have language-wide validity can be successfully combined with information at superficial or semantic level.

In this paper, we present an integration of two kinds of discursive information, *cohesion* and *coherence*, to obtain an adequate representation of text for the task of TS. Our starting point is an extractive informative summarization system that exploits the cohesive properties of text by building and ranking lexical chains (see Section 3). This system is enhanced with discourse coherence information (Section 5.3). Experiments were carried out on the combination of these two kinds of information, and results were evaluated on a Spanish news agency corpus (Section 5).

2 Previous Work on Combining Cohesion and Coherence

Traditionally, two main components have been distinguished in the discursive structure of a 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.* (Mani, 2001) argues that an integration of these two kinds of discursive information would yield significant improvements in the task of text summarization.

(Corston-Oliver and Dolan, 1999) showed that eliminating discursive satellites as defined by the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), yields an improvement in the task of Information Retrieval. Precision is improved because only words in discursively relevant text locations are taken into account as indexing terms, while traditional methods treat texts as unstructured bags of words.

Some analogous experiments have been carried out in the area of TS. (Brunn et al., 2001; Alonso and Fuentes, 2002) claim that the performance of summarizers based on lexical chains can be improved by ignoring possible chain members if they occur in irrelevant locations such as subordinate clauses, and therefore only consider chain candidates in main clauses. However, syntactical subordination does not always map discursive relevance. For example, in clauses expressing finality or dominated by a verb of cognition, like *Y said that X*, the syntactically subordinate clause *X* is discursively nuclear, while the main clause is less relevant (Verhagen, 2001).

In (Alonso and Fuentes, 2002), we showed that identifying and removing discursively motivated satellites yields an improvement in the task of text summarization. Nevertheless, we will show that a more adequate representation of the source text can be obtained by ranking chain members in accordance to their position in the discourse structure, instead of simply eliminating them.

3 Summarizing with Lexical Chains

The lexical chain summarizer follows the work of (Morris and Hirst, 1991) and (Barzilay, 1997).

As can be seen in Figure 1 (left) the text is first segmented, at different granularity levels (paragraph, sentence, clause) depending on the application. To detect chain candidates, the text is morphologically analysed, and the lemma and POS of each word are obtained. Then, Named Entities are identified and classified in a gazzetteer. For Spanish, a simplified version of (Palomar et al., 2001) extracts co-referenece links for some types of pronouns, dropping off the constraints and rules involving syntactic information.

Semantic tagging of common nouns is been performed with *is-a* relations by attaching EuroWordNet (Vossen, 1998) synsets to them. Named Entities are been semantically tagged with *instance* relations by a set of *trigger words*, like *former president*, *queen*, etc., associated to each of them in a gazzetteer. Semantic relations between common nouns and Named Entities can be established via the EWN synset of the trigger words associated to a each entity.

Chain candidates are common nouns, Named Entities, definite noun phrases and pronouns, with no word sense disambiguation. For each chain candidate, three kinds of relations are considered, as defined by (Barzilay, 1997):

- **Extra-strong** between repetitions of a word.
- **Strong** between two words connected by a direct EuroWordNet relation.
- **Medium-strong** if the path length between the EuroWordNet synsets of the words is longer than one.

Being based on general resources and principles, the system is highly parametrisable. It has a relative independence because it may obtain summaries for texts in any language for which there is a version of WordNet an tools for POS tagging and Named Entity recognition and classification. It can also be parametrised for obtaining summaries of various lengths and at granularity levels.

As for relevance assessment, some constraints can be set on chain building, like determining the maximum distance between WN synsets of chain candidates for building medium-strong chains, or the type of chain merging when using gazetteer information. Once lexical chains are built, they are scored according to a number of heuristics that consider characteristics such as their length, the kind of relation between their words and the point of text where they start. Textual Units (TUs) are ranked according to the number and type of chains crossing them, and the TUs which are ranked highest are extracted as a summary. This ranking of TUs can be parametrised so that a TU can be assigned a different relative scoring if it is crossed by a strong chain, by a Named Entity Chain or by a co-reference chain. For a better adaptation to textual genres, heuristics schemata can be applied.

However, linguistic structure is not taken into account for scoring the relevance lexical chains or TUs, since the relevance of chain elements is calculated irrespective of other discourse information. Consequently, the strength of lexical chains is exclusively based on lexic. This partial representation can be even misleading to discover the relevant elements of a text. For example, a Named Entity that is nominally conveying a piece of news in a document can present a very tight pattern of occurrence, without being actually relevant to the aim of the text. The same applies to other linguistic structures, such as recurring parallelisms, examples or adjuncts. Nevertheless, the relative relevance of these elements is usually marked structurally, either by sentential or discursive syntax.

4 Incorporating Rhetorical and Argumentative Relations

The lexical chain summarizer was enhanced with discourse structural information as can be seen in Figure 1 (right).

Following the approach of (Marcu, 1997), a partial representation of discourse structre was obtained by means of the information associated to a Discourse Marker (DM) lexicon. DMs are described in four dimensions:

- **matter**: following (Asher and Lascarides, 2002), three different kinds of subject-matter meaning are distinguished, namely *causality*, *parallelism* and *context*.
- **argumentation**: in the line of (Anscombre and Ducrot, 1983), three argumentative moves are distinguished: *progression*, *elaboration* and *revision.*
- **structure**: following the notion of right frontier (Polanyi, 1988), *symmetric* and *asymmetric* relations are distinguished.
- **syntax**: describes the relation of the DM with the rest of the elements at the discourse level, in the line of (Forbes et al., 2003), mainly used for discourse segmentation.

The information stored in this DM lexicon was used for identifying inter- and intra-sentential discourse segments (Alonso and Castellón, 2001) and the discursive relations holding between them. Discourse segments were taken as Textual Units by the Lexical Chain summarizer, thus allowing a finer granularity level than sentences.

Two combinations of DM descriptive features were used, in order to account for the interaction of different structural information with the lexical information of lexical chains. On the one hand, *nucleus-satellite* relations were identified by the combination of *matter* and *structure* dimensions of DMs. This **rhetorical** information yielded a hierarchical structure of text, so that satellites are subordinate to nucleus and they are accordingly considered less relevant. On the other hand, the **argumentative** line of text was traced via the *argumentation* and also *structure* DM dimensions, so that segments were tagged with their contribution to the progression of the argumentation.

These two kinds of structural analyses are complementary. Rhetorical information is mainly effective at discovering local coherence structures, but it is unreliable when analyzing macrostructure. As (Knott et al., 2001) argue, a different kind of analysis is needed to track coherence throughout a whole text; in their case the alternative information used is focus, we have opted for argumentative orientation. Argumentative information accounts for a higher-level structure, although it doesn't provide much detail about it.

This lexicon has been developed for Spanish (Alonso et al., 2002a). Nevertheless, the structure of the DM lexicon and the discourse parsing tools based on it is highly portable, and versions

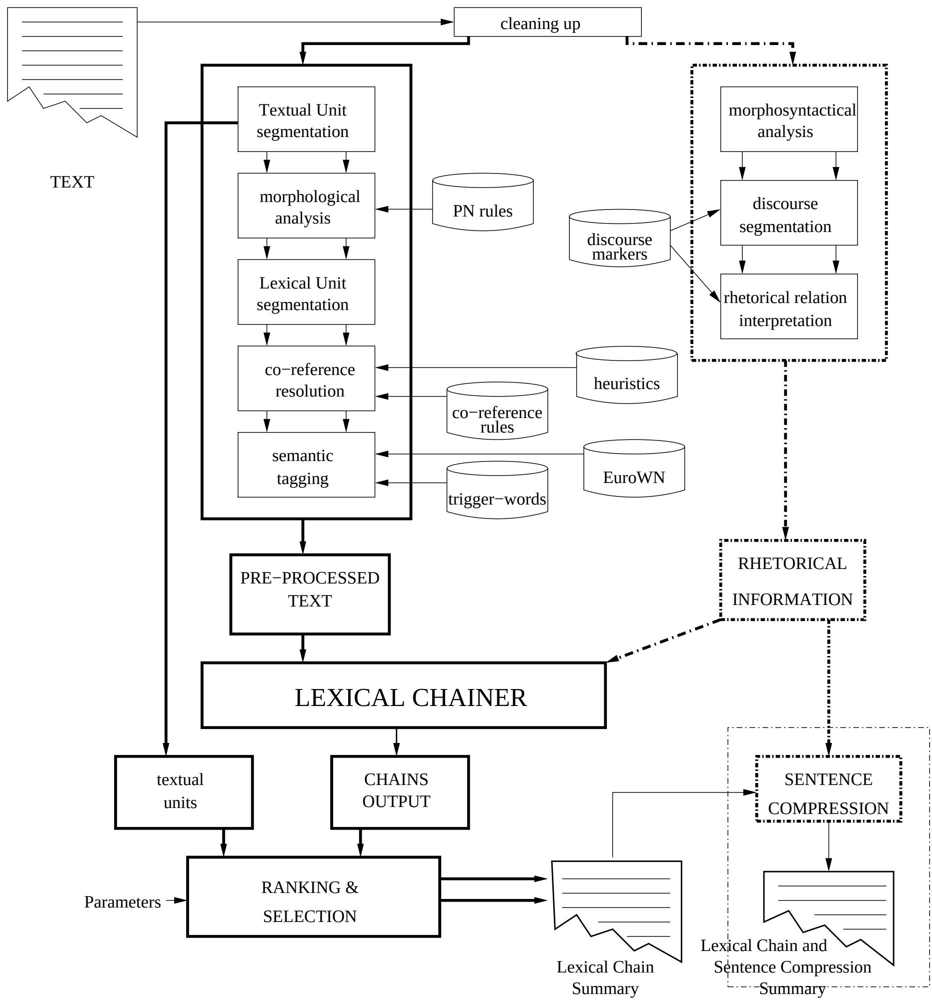

Figure 1: Integration of discursive information: lexical chains (left) and discourse structural (right)

for English and Catalan are being developed by bootstraping techniques (Alonso et al., 2002b).

5 Experiments

A number of experiments were carried out in order to test whether taking into account the structural status of the textual unit where a chain member occurs can improve the relevance assessment of lexical chains (see Figure 2). Since the DM lexicon and the evaluation corpus were available only for Spanish, the experiments were limited to that language. Linguistic pre-processing was performed with the CLiC-TALP system (Carmona et al., 1998; Arévalo et al., 2002).

For the evaluation of the different experiments, the evaluation software MEADeval (MEA, 2002) was used, to compare the obtained summaries with a golden standard (see Section 5.1). From this package, the usual precision and recall measures were selected, as well as the simple cosine. Simple cosine (simply *cosine* from now on) was chosen because it provides a measure of similarity between the golden standard and the obtained extracts, overcoming the limitations of measures depending on concrete textual units.

5.1 Golden Standard

The corpus used for evaluation was created within Hermes project[1], to evaluate automatic summariz-

[1] Information about this project available in http://terral.ieec.uned.es/hermes/

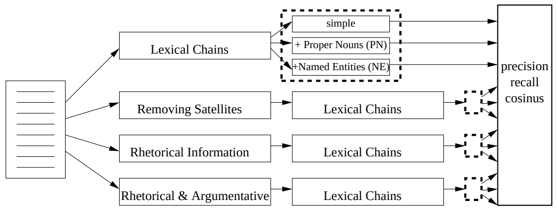

Figure 2: Experiments to assess the impact of discourse structure on lexical chain members

ers for Spanish, by comparison to human summarizers. It consists of 120[2] news agency stories of various topics, ranging from 2 to 28 sentences and from 28 to 734 words in length, with an average length of 275 words per story.

To avoid the variability of human generated abstracts, human summarizers built an extract-based golden standard. Paragraphs were chosen as the basic textual unit because they are self-contained meaning units. In most of the cases, paragraphs contained a single sentence. Every paragraph in a story was ranked from 0 to 2, according to its relevance. 31 human judges summarized the corpus, so that at least 5 different evaluations were obtained for each story.

Golden standards were obtained coming as close as possible to the 10% of the length of the original text (19% compression average).

The two main shortcomings of this corpus are its small size and the fact that it belongs to the journalistic genre. However, we know of no other corpus for summary evaluation in Spanish.

5.2 Performance of the Lexical Chain System

The performance of the Lexical Chain System with no discourse structural information was taken as the base to improve. (Fuentes and Rodríguez, 2002) report on a number of experiments to evaluate the effect of different parameters on the results of lexical chains. To keep comparability with the golden standard, and to adequately calculate precision and recall measures, paragraph-sized TUs were extracted at 10% compression rate.

Some parameters were left unaltered for the whole of the experiment set: only *strong* or *extra-strong* chains were built, no information from defined noun phrases or trigger words could be used and only short co-reference chains were built. Results are presented in Table 1.

[2]For the experiments reported here, one-paragraph news were dropped, resulting in a final set of 111 news stories.

	Precision	Recall	Cosine
Lead	.95	.85	.90
SweSum	.90	.81	.87
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
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

Table 1: Performance of the lexical chain Summarizer

The first column in the table shows the main parameters governing each trial: simple lexical chains, lexical chains successively augmented with proper noun and co-Reference chains, and finally giving special weighting to the 1st TU because of global document structure appliable to the journalistic genre.

Two heuristics schemata were experimented: *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. An evaluation of SweSum (SweSum, 2002), a summarization system available for Spanish, is also provided as a comparison ground. Trials with SweSum were carried out with the default parameters of the system. In addition, the first paragraph of every text, the so-called lead summary, was taken as a dummy baseline.

As can be seen in Table 1, the lead achieves the best results, with almost the best possible score. This is due to the pyramidal organisation 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 and *heuristic 1*.

However, it must be noted that lexical chain summarizer produces results with high cosine and low precision, while SweSum yields high precision and low cosine. This means that, while the textual units extracted by the summarizer are not identical to the ones in the golden standard, their content is not dissimilar. This seems to indicate that the summarizer successfully captures content-based relevance, which is genre-independent. Consequently, the lexical chain summarizer should be able to capture relevance when applied to non-journalistic texts. This seems to be supported by the fact that *heuristic 2* improves cosine over precision four points higher than *heuristic 1*, which seems more genre-dependent.

Unexpectedly, co-reference chains cause a decrease in the performance of the system. This may be due to their limited length, and also to the fact that both full forms and pronouns are given the same score, which does not capture the difference in relevance signalled by the difference in form.

	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 2: Results of the integration of lexical chains and discourse structural information

5.3 Results of the Integration of Heterogenous Discursive Informations

Structural discursive information was integrated with only those parameters of the lexical chain summarizer that exploited general discursive information. *Heuristic 1* was not considered because it is too genre-dependent. No co-reference information was taken into account, since it does not seem to yield any improvement.

The results of integrating lexical chains with discourse structural information can be seen in Table 2. Following the design sketched in Figure 5, the performance of the lexical chains summarizer was first evaluated on a text where satellites had been removed. As stated by (Brunn et al., 2001; Alonso and Fuentes, 2002), removing satellites slightly improves the relevance assessment of the lexical chainer (by one point).

Secondly, discourse coherence information was incorporated. Rhetorical and argumentative informations were distinguished, since the first identifies mainly unimportant parts of text and the second identifies both important and unimportant. 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 .5 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. When this parameter is set, removing satellites outperforms the results obtained by taking into account discourse structural information in precision. However, this can also be due to the fact that when the text is compressed, TUs are shorter, and a higher number of them can be extracted within the fixed compression rate. It must be noted, though, that recall does not drop for these summaries.

Lastly, intra-sentential and 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 summaries to a 14.43% for summaries which were sentence-compressed. Moreover, since sentences were shortened, readability was increased, which can be considered as a further factor of compression. However, these summaries have not been evaluated with the MEADeval package because no golden standard was available for textual units smaller than paragraphs. Precision and recall measures could not be calculated for summaries that removed satellites, because they could not be compared with the golden standard, consisting only full sentences.

5.4 Discussion

The presented evaluation successfully shows the improvements of integrating cohesion and coherence, but it has two weak points. First, the small size of the corpus and the fact that it represents a single genre, which does not allow for safe generalisations. Second, the fact that evaluation metrics fall short in assessing the improvements yielded by the combination of these two discursive informations, since they cannot account for quantitative improvements at granularity levels different from the unit used in the golden standard, and therefore a full evaluation of summaries involving sentence compression is precluded. Moreover, qualitative improvements on general text coherence cannot be captured, nor their impact on summary readability.

As stated by (Goldstein et al., 1999), "*one of the unresolved problems in summarization evaluation is how to penalize extraneous non-useful information contained in a summary*". We have tried to address this problem by identifying text segments which carry non-useful information, but the presented metrics do not capture this improvement.

6 Conclusions and Future Work

We have shown that the collaborative integration of heterogeneous discursive information yields an improvement on the reperesentation of source text, as can be seen by improvements in resulting summaries. Although this enriched representation does not outperform a dummy baseline consisting of taking the first paragraph of the text, we have argued that the resulting representation of text is genre-independent and succeeds in capturing content relevance, as shown by cosine measures.

Since the properties exploited by the presented system are text-bound and follow general principles of text organization, they can be considered to have language-wide validity. This means that the system is domain-independent, though it can be easily tuned to different genres.

Moreover, the system presents portability to a variety of languages, as long as it has the knowledge sources required, basically, shallow tools for morpho-syntactical analysis, a version of WordNet for building and ranking lexical chains, and a lexicon of discourse markers for obtaining a certain discourse structure.

Future work concerning the lexical chain summarizer will be focussed in building longer lexical chains, exploiting other relations in EWN, merging chains and even merging heterogeneous information. Improvements in the analysis of structural discursive information include enhancing the scope to paragraph and global document level, integrating heterogeneous discursive information and proving language-wide validity of Discourse Marker information.

To provide an adequate assessment of the achieved improvements, the evaluation procedure is currently being changed. Given the enormous cost of building a comprehensive corpus for summary evaluation, the system has been partially adapted to English, so that it can be evaluated with the data and procedures of (DUC, 2002).

Nevertheless, our future efforts will also be directed to gathering a corpus of Spanish texts with abstracts from which to automatically obtain a corpus of extracts with their corresponding texts, as proposed by (Marcu, 1999). Concerning qualitative evaluation, we will try to apply evaluation metrics that are able to capture content and coherence aspects of summaries, such as more complex content similarity or readability measures.

7 Acknowledgements

This research has been conducted thanks to a grant associated to the X-TRACT project, PB98-1226 of the Spanish Research Department. It has also been partially funded by projects HERMES (TIC2000-0335-C03-02), PETRA (TIC2000-1735-C02-02), and by CLiC (Centre de Llengutatge i Computació).

References

Laura Alonso and Irene Castellón. 2001. Towards a delimitation of discursive segment for natural language processing applications. In *First International Workshop on Semantics, Pragmatics and Rhetoric*, Donostia - San Sebastiàn, November.

Laura Alonso and Maria Fuentes. 2002. Collaborating discourse for text summarisation. In *Proceedings of the Seventh ESSLLI Student Session*.

Laura Alonso, Irene Castellón, and Lluís Padró. 2002a. Design and implementation of a spanish discourse marker lexicon. In *SEPLN*, Valladolid.

Laura Alonso, Irene Castellón, and Lluís Padró. 2002b. X-tractor: A tool for extracting discourse markers. In *LREC 2002 workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*, Las Palmas.

J. C. Anscombre and O. Ducrot. 1983. *L'argumentation dans la langue*. Mardaga.

Montse Arévalo, Xavi Carreras, Lluís Màrquez, M.Antònia Martí, Lluís Padró, and M.José Simón. 2002. A proposal for wide-coverage spanish named entity recognition. *Procesamiento del Lenguaje Natural*, 1(3).

Nicholas Asher and Alex Lascarides. 2002. *The Logic of Conversation*. Cambridge University Press.

Regina Barzilay. 1997. *Lexical Chains for Summarization*. Ph.D. thesis, Ben-Gurion University of the Negev.

Meru Brunn, Yllias Chali, and Christopher J. Pinchak. 2001. Text Summarization using lexical chains. In *Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001*, New Orleans, Louisiana.

Josep Carmona, Sergi Cervell, Lluís Màrquez, M. Antònia Mart, Lluís Padró, Roberto Placer, Horacio Rodríguez, Mariona Taulé, and Jordi Turmo. 1998. An environment for morphosyntactic processing of unrestricted spanish text. In *First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain.

Simon H. Corston-Oliver and W. Dolan. 1999. Less is more: Eliminating index terms from subordinate clauses. In *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 348 – 356.

DUC. 2002. DUC–document understanding conference. http://duc.nist.gov/.

K. Forbes, E. Miltsakaki, R. Prasad, A. Sarkar, A. Joshi, and B. Webber. 2003. D-LTAG system - discourse parsing with a lexicalized tree-adjoining grammar. *Journal of Language, Logic and Information*. to appear.

Maria Fuentes and Horacio Rodríguez. 2002. Using cohesive properties of text for automatic summarization. In *JOTRI'02*.

Jade Goldstein, Vibhu Mittal, Mark Kantrowitz, and Jaime Carbonell. 1999. Summarizing text documents: Sentence selection and evaluation metrics. In *SIGIR-99*.

M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. English Language Series. Longman Group Ltd.

Alistair Knott, Jon Oberlander, Mick O'Donnell, and Chris Mellish. 2001. Beyond elaboration: The interaction of relations and focus in coherent text. In Ted Sanders, Joost Schilperoord, and Wilbert Spooren, editors, *Text representation: linguistic and psycholinguistic aspects*, pages 181–196. Benjamins.

Inderjeet Mani. 2001. *Automatic Summarization*. Nautral Language Processing. John Benjamins Publishing Company.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organisation. *Text*, 3(8):234–281.

Daniel Marcu. 1997. *The Rhetorical Parsing, Summarization and Generation of Natural Language Texts*. Ph.D. thesis, Department of Computer Science, University of Toronto, Toronto, Canada.

Daniel Marcu. 1999. The automatic construction of large-scale corpora for summarization research. In *SIGIR-99*.

2002. MEADeval. http://perun.si.umich.edu/clair/meadeval/.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion, the thesaurus, and the structure of text. *Computational linguistics*, 17(1):21–48.

M. Palomar, A. Ferrández, L. Moreno, P. Martínez-Barco, J. Peral, M. Saiz-Noeda, and R. Mu noz. 2001. An algorithm for anaphora resolution in spanish texts. *Computational Linguistics*, 27(4).

Livia Polanyi. 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638.

R. Schank and R. Abelson. 1977. *Scripts, Plans, Goals, and Understanding*. Lawrence Erlbaum, Hillsdale, NJ.

SweSum. 2002. http://www.nada.kth.se/˜ xmartin/swesum/index-eng.html.

Arie Verhagen. 2001. Subordination and discourse segmentation revisited, or: Why matrix clauses may be more dependent than complements. In Ted Sanders, Joost Schilperoord, and Wilbert Spooren, editors, *Text Representation. Linguistic and psychological aspects*, pages 337–357. John Benjamins.

Piek Vossen, editor. 1998. *Euro WordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers.