Collaborating Discourse for Text Summarisation

Laura Alonso Alemany
CLiC, Department of General Linguistics, 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. Traditionally, two main lines have been distinguished in discourse processing: *cohesion-based* and *coherence-based*. They are strongly correlated, but they have worked independently in NLP applications. We present how a cohesion-based method and a coherence-based one can successfully collaborate in obtaining an enriched representation of discourse that is useful for knowledge requiring NLP task, Automated Text Summarisation (TS).

1 Introduction

Obtaining the discourse structure of a text is useful for a variety of NLP tasks, especially knowledge-requiring ones such as natural language generation, machine translation or Automated Text Summarisation (TS).

Traditionally, two main components have been distinguished in the discursive structure of a source text: cohesion and coherence. As defined by Halliday y 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.

As follows, cohesion and coherence account for complementary aspects of the discourse structure of a text. Some theoretical approaches to discourse processing, such as Polanyi (1988), give an integrated account of a number of linguistic levels, including these two. But, however evident this

Proceedings of the Seventh ESSLLI Student Session
Malvina Nissim (editor)

integration may seem from a theoretical point of view, cohesion-based discourse models and coherence-based ones have usually worked separatedly for general-purpose NLP applications, probably because the complexity of dealing with both is unaffordable for current systems.

In this paper, we will present an initial collaboration between two such complementary approaches, Lexical Chains and Rhetorical Structure. While the first account for the cohesive aspects of the source text, the second captures its coherence structure. Task-based evaluation seems an adequate method to assess the improvements in the representation of discourse that the collaboration of these two methods may yield. Therefore, we have chosen to carry out an evaluation within the framework of a knowledge-requiring NLP task, namely discourse-based TS.

The rest of the paper is as follows: in Section 2 we will present the partial solutions that these two techniques provide for the TS problem separatedly, showing how the strengths of each are complementary of the shortcomings of the other. In Section 3 we will present the systems that implement Lexical Chains and Rhetorical Structure Analysis for TS in Spanish. Section 4 outlines the experiment by which we have tried to assess the improvements of the collaboration of these two methods. We will finish with the discussion of the obtained results and future work (Section 5).

2 Discourse Structure for Text Summarisation

As defined by Sparck-Jones (2001), "*A summary is a reductive transformation of a source text into a summary text by extraction or generation*". Therefore, the problem of TS can be divided into:

1. **Locating** the relevant fragments, i.e. sentences, paragraphs, passages, of a source text
2. **Ranking** these fragments by relevance and
3. **Producing** a summary, either an *extract*, if the summary is composed by literal fragments of text, or an *abstract*, if it is generated

A number of techniques have been applied to address these points, going from shallow statistical processing with no linguistic knowledge to deep understanding of texts. Some of these techniques are: lexical chains, (Barzilay, 1997), coreference chains, (Baldwin y Morton, 1998), alignment techniques, (Banko et al., 1999), similarity and divergence measures, (Carbonell y Goldstein, 1998), statistical models, (Schlesinger, Baker, y Donaway, 2001), (Conroy et al., 2001), Machine Learning approaches, (Knight y Marcu, 2000), (Tzoukermann, Muresan, y Klavans, 2001), sentence reduction, (Jing, 2001), Information Extraction techniques, (Kan y McKeown, 1999), topic detection-based systems, as (Hovy y Lin, 1999), or systems using the rhetorical structure of the document, (Marcu, 1997a). Current approaches to TS try to merge heterogeneous textual information to obtain an enriched representation of the source text, by combining some of these techniques, as in (Kraaij, Spitters, y van der Heijden, 2001), (Muresan, Tzoukermann, y Klavans, 2001), (White et al., 2001).

Discourse-based TS is aimed at obtaining a certain representation of the discursive dimension of the source text. The underlying assumption of this approach is that a more adequate summary can be obtained from a representation of the source text that captures its discourse structure. By analysing discourse structure, a better account of relevance can be achieved, and also improvements in readability.

As in most NLP tasks, two main lines can be distinguished in discourse-based TS. Among cohesion-based methods, lexical chains (see Section 2.1) has been widely applied, because it is highly portable, relying on general lexical properties. It is able to reflect a structured representation of the content of a text by expressing cohesion ties between its words. On the other hand, methods that apply a shallow approach to coherence-based discourse summarisation have become very popular since the mid-nineties (see Section 2.2). Applications of RST (Mann y Thompson, 1988) are able to capture relevance and coherence relations between parts of text with no need for costly and domain-dependent resources.

When trying to achieve a representation of text that permits locating and ranking relevant fragments, these two approaches prove to be highly complementary. Lexical chains account for the linear *distribution of content* in a text, considering as most relevant those fragments of text where most of the identified content lines are represented. On the other hand, coherence relations provide a *hierarchical strucutre* of the same text, with most relevant fragments placed at the top of the hierarchy. Each of these methods yields a partial, but mutually complementary, representation of discourse.

However complementary these two approaches might be, there is a lack for real collaboration in their application to TS. (Brunn, Chali, y Pinchak, 2001) propose the use of a lexical-chain based summarizer that incorporates shallow discourse marker information. In this initial proposal there is no real interaction between the two kinds of information, because they are applied sequentally. Our initial hypothesis is that a better representation of discourse can be achieved by combining these two kinds of information. The improvement yielded by the combination of these two techniques can be assessed by an improvement in the task of TS.

2.1 Lexical Chains

Lexical Chains try to identify cohesion links between parts of text by identifying relations holding between their words. Two pieces of text are considered to be lexically related not only if they use the same words, but also if they use semantically related words. This is a way to obtain a certain structure of a text based on the distribution of its content.

Hasan (1984) establishes that *identity chains* contain terms that refer to the same object. They are created by pronominal cohesion, lexical repetition or instantial equivalence and are always text-bound, because the relation of co-reference can be determined only in the context of a text. In contrast, *similarity chains* are not text-bound. Their elements are held together by semantic bonds obtained throuhgh a lexical resource. These bonds are supra-textual, with a language-wide validity. The two types of chains are important for cohesion analysis, however, the advantages of similarity chains over identity is that their implementation is simpler, since they can be computed with no deep text understanding.

Lexical Chains provide a representation of text that has been used for a variety of NLP tasks, including topic passage segmentation (Hearst, 1994), topic detection (Lin y Hovy, 1997) detection of malapropisms (Hirst y St-Onge, 1997), automated text summarisation (Barzilay, 1997) or automatic hypertext generation (Green, 1997). The general procedure for constructing lexical chains follows three steps:

1. Select a set of candidate words
2. For each candidate word, find an appropriate chain relying on a *relatedness criterion* among members of the chains. Relatedness of words is determined in terms of the distance between their occurrences and the shape of the path connecting them in WordNet.
3. If a chain is found, insert the word in the chain and update it accordingly.

Chains are scored according to a number of heuristics: their length, the kind of relation between their words, the point of text where they start, etc.

One of the drawbacks of lexical chains is that they are insensitive as to the non-lexical structure of texts, 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, sometimes not even within the layout- or genre-determined structure of the document. It can be said, then, that lexical chains rely exclusively on paradigmatic relations between chain members: identity, co-reference, WN-typological relations, etc. Only proximity relations take into account the contextual aspect of chain members. This is descriptively inadequate, since some syntagmatic relations, like syntactic role or discursive salience, can significantly determine the relevance of an occurrence of a chain member, and contribute to shape the resulting chain.

2.2 Rhetorical Structure

The classical approaches to obtaining a representation of textual coherence heavily rely on huge sources of hand-coded knowledge that enable systems to carry out the inferences to derive the coherence structure of a text. Due to their high computational cost, applications of these approaches have just worked for small systems in restricted domains.

Some approaches to summarisation developed since the mid 1990s supposed a turning point to the way textual coherence was exploited for NLP. Good examples are Marcu (1997b) and Ono, Sumita, y Miike (1994), who apply the Rhetorical Structure Theory (RST) proposed by Mann y Thompson (1988). They obtain the rhetorical structure of a text by exploiting surface coherence clues, such as punctuation or discourse markers. This clues serve to identify discourse units and the relationship between them, in the form of a hierarchical structure that reflects the relative importance of every part of the text.

Nevertheless, as the authors themselves state, shallow coherence clues are often insufficient to automatically build a structure of text that is informative enough to obtain a reliable summary. While they seem to give an accurate account of discourse microstructure, they are often too ambiguous as to the macrostructural level. This is mainly due to the limited scope of discourse markers as unambiguous discourse operators, and also to the fact that content-based relations, such as *elaboration*, cannot be captured. Moreover, linear aspects of discourse tend to be misrepresented in a hierarchical representation such as that proposed by RST. Satisfaction-precedence relations, like reference to entities, events or even arguments, are not adequately handled, although they consitute an important source of textuality (Grosz y Sidner, 1986). All this tends to result in a loss in coherence at a macro-structural level.

Knott et al. (2001) argue that an improved representation of discourse can be obtained by combining hierarchical and linear discursive information. They propose to organise text as a sequence of focus-spaces occupied by small RST-like trees that represent chunks of discourse with a tight and clear coherence structure. The top of each of these trees is the entity in focus. We believe this is a good example of the complementary function of coherence and cohesive devices in a comprehensive account of discourse *organisation.

3 Proposed Summarisation Systems

We are going to evaluate the interaction between two discourse-based summarisation systems for Spanish, one applying a Lexical Chain approach and another that exploits Rhetorical Structure. Lexical chains can produce extracts at different granularity levels and compression rates, while rhetorical analysis granularity is determined by the structure of the text. Both of them rely on the morphosyntactical analysis given by the CLiC-TALP system (Carmona et al., 1998; Padró, 1997; Atserias, Castellón, y Civit, 1998). The full architecture of the proposed systems is described below.

3.1 Summarisation by Lexical Chains

The Lexical Chain system is the one proposed in Fuentes y Rodríguez (2002). It follows the work of Barzilay (1997). First, the text is segmented, with varying degrees of granularity depending on the application. To detect chain candidates, the text is pre-processed, and the lemma and POS of each word are obtained. Additionally, the information stored in a named entity Gazeteer is associated to named entities. This Gazetteer is based in a MUC-like ontology, with trigger words associated to each entity. In turn, these trigger words are linked to EuroWordNet synsets. Chain candidates are common nouns, proper nouns, named entities, definite noun phrases and pronouns. For each candidate word, three kinds of relations are considered, as defined by Hirst y St-Onge (1997):

- **Extra-strong** between a word and its repetition
- **Strong** between two words connected by a EuroWordNet[1] relation
- **Medium-strong** if the link between the EuroWordNet synsets of the words is longer than one.

In addition, this system establishes relations between common nouns and the rest of chain candidates (proper nouns and named entities), by way of the information stored in the Gazetteer.

To build the chains, there are constraints on the path length according the type of edges, determined by the relations established between chain members.

Chains are scored so that strong chains are identified. Sentences are ranked and those crossed by most strong chains are considered to be most relevant. A certain number of sentences is extracted from this ranked list until a determined summary length is achieved.

3.2 Summarisation by Rhetorical Structure

The system that exploits Discourse Markers is the one proposed in Alonso (2001). It works on a morphosyntactically tagged text and consists of three modules:

[1]Relations between words are computed using the Spanish version of EuroWordNet instead of WordNet.

1. **Segmenter**, it identifies unambiguous minimal discursive segments via surface clues such as punctuation, syntactical chunks or discourse markers.
2. **Interpreter**, it identifies coherence relations holding between minimal discursive segments by the kind of segment or the presence of a discourse marker.
3. **Discourse Marker Lexicon**, containing 600 cue phrases.

The outcome of this system is designed to give a representation of coherence relations holding at three levels: sentence, informative block (paragraph-like) and full text. Coherence relations between segments are informative as to the relative relevance of each segment in respect to other segments in the same level, so that only the most relevant are included in the final summary. Also segments with a co-dependence relation with a relevant segment are included.

4 Experiment

The aim of this experiment is to show whether the presented coherence-based and cohesion-based approaches to discourse for text summarisation can meaningfully collaborate and if this collaboration improves the resulting representation of discourse.

For this first integration of the two approaches, we have implemented the simplest version of each system. Accordingly, lexical chains are built only considering *extra-strong relations* between words, identity or WN-synonymy relations. Rhetorical relations are established only at the microstructural level, within a sentence scope. The results of this first approach will direct a richer collaboration of the two lines in the future.

To assess the improvements achieved, the extracts obtained by combination of the two systems were compared against the performance of each system separatedly. For further comparison, we took two baselines: a lead-based summary, that takes the first 10% of the text, and a frequency-based summary, that applies the *tf*idf* method on a shallow lemmatized text to extract the sentences containing the most frequent words. The whole experiment design is in Figure 1.1.

To evaluate automatically obtained summaries, a number of methods have been proposed (Baldwin et al., 2000; Radev, Jing, y Budzikowska, 2000; McKeown et al., 2001). Usually, a distinction is made between extrinsic, task-based evaluation and intrinsic evaluation, taking into account the quality of the summary by itself or by comparison with a *golden standard*. A *golden standard* is the ideal summary that the system is intended to produce.

We worked on a corpus of 111 news agency stories of various topics from the Hermes project[2]. To avoid the variability of human generated abstracts, an extract-based golden standard was created from this corpus. The textual units to be extracted were paragraph-sized, since they were found to be natural meaning units for human judges. Judges ranked every paragraph in a story from 0 to 2, according to its relevance. Each summary in the golden standard was then obtained by selecting the highest ranked paragraphs in the story, coming as close as possible to the 10% of the length of the original text (with an average 19% compression).

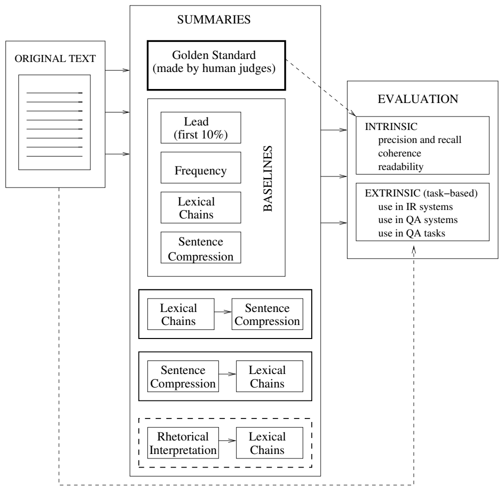

Figure 1.1: Experiment Design

Automatic summaries were compared with this golden standard at paragraph level, by precision and recall metrics. However, it was not possible to obtain these metrics for rhetorical structure summaries, because this system operates with textual units smaller than paragraphs. Moreover, their compression rate, more than 50%, is not comparable with 10% golden standards.

5 Results

The results of our expermiment showed an improvement on the representation of texts when the two kinds of discursive information are taken into account. Precision and recall figures in Table 1.1 confirm an improvement on the resulting summaries. Also compression was increased, by an approximate 20%. The best results are obtained when the summarising process comes closer to a deep interaction of the two kinds of information: to have the texts rhetorically interpreted and unimportant segments at sentence level removed (sentence compression), and then apply lexical chains onto these relevance-pruned texts.

When compared with baseline summaries, we found that any system performs better than the frequency baseline. However, it must be taken into account that the textual units of this summarizer were sentence-sized, so comparison with the golden standard was carried out by considering the paragraph where the extracted sentences occurred. Therefore, the presented precision figures might be lower than the actual performance of this system, while recall might be higher.

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

	Precision	Recall
Baselines		
Lead	.92	.86
Frequency	.44	.56
Single Methods		
Lexical Chains	.73	.72
Sentence Compression	not appl.	not appl.
Integration		
Lexical Chains + Sentence Compression	.73	.72
Sentence Compression + Lexical Chains	.77	.71

Table 1.1: Comparison of Summarisation Methods at 19% compression rate

Not surprisingly, the simplest method, the lead, obtains higher precision and recall figures than any other method, being overriden only as to compression by the combination of lexical chains and sentence compression. This is very common in short articles of journalistic genres, because journalistic style determines that information is ordered by relevance, so that more important information is placed first in the article (the so-called *pyramidal structure* of articles). However, this structure is not extrapolable to long texts or different genres.

In Figure 1.2 we can see an example of how an inadequate discourse representation by lexical chains or rhetorical structure alone can be improved by collaboration. In this case, lexical chains (left) signal as the most relevant part of text the one crossed by most strong chains (in the thickest box). But the most important part of this text is the first paragraph. As for rhetorical structure (right), the most important part of the text cannot be safely determined, because discourse markers only provide unambiguous local relevance assessment. However, if the non-relevant parts of text signalled by rhetorical structure are first removed, the previously selected paragraph is not available any more for extraction. The most relevant part of text will be now the first paragraph, which is the same as in the golden standard.

For this experiment, the two systems worked only sequentially, with no real interactive combination of the two kinds of information (in the dotted box in Figure 1.1). A deeper interaction would imply that lexical chains become sensitive to the hierarchichal structure of texts, for example, if every chain candidate provides a relative score to the chain according to the relative relevance of the segment where it occurs. In the above example, inadequate strong lexical chains would be weakened, according to the number of chain members occurring in irrelevant text segments. Also rhetorical interpretation could benefit from lexical chain information; indeed, we found that 52.2% of the errors produced in analysis of the rhetorical structure could be solved with lexical chain information.

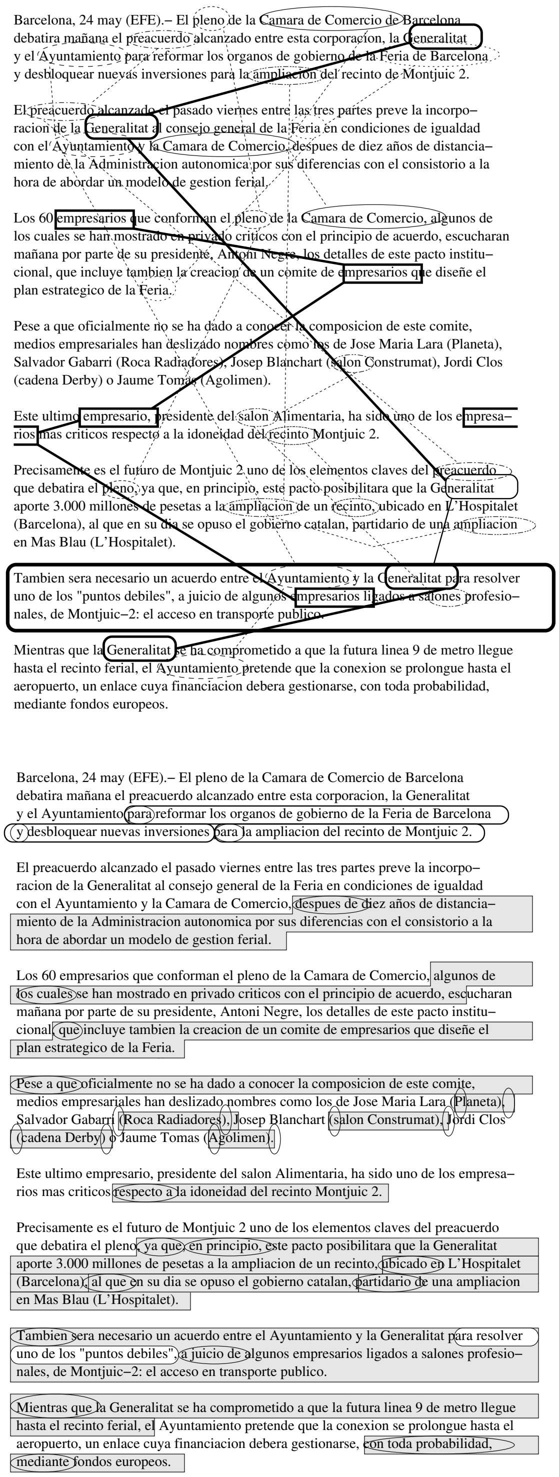

Figure 1.2: A text represented by lexical chains (left) and rhetorical structure (right). Boxes represent chain members, lines are chains, thick ones being strong chains. Shadowed boxes contain rhetorically subordinate text segments, white boxes contain central segments and discourse markers are in circles.

5.1 Evaluation shortcomings

The most evident shortcoming of the presented evaluation is the small size of the evaluation corpus. This is a consequence of the high cost of building evaluation resources that require human judgement. In addition, the granularity of the golden standard supposed a further restriction on the granularity of the summaries that could be evaluated. For example, *lexical chains + sentence compression* summaries have the same precision and recall figures that *lexical chains* summaries, because sentence compression summarizer operates at intra-sentential level, which cannot be captured by precision and recall metrics at paragraph level. 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*".

In general terms, it must be said that there is a lack for adequate evaluation metrics for summaries. Alternative methods are proposed by Donaway, Drummey, y Mather (2000) (content similarity) and Radev, Jing, y Budzikowska (2000) (utility method). These methods are not dependant on units, so they are applicable to summaries of different granularities. However, they focus in content representativity, the relative relevance of lexic elements in respect to the structure of discourse is still not captured.

Another unresolved problem is the assessment of summary coherence, which is in general limited to the grammaticality of the sentences and to resolution of entity anaphoras. Other aspects that clearly influence summary readability are not taken into account, such as sentence length, sentence complexity, or complex referential coherence (event or argument anaphora). The integration of the two kinds of information that we have presented here tries to capture some of these aspects of textual coherence and offer a simplified summary text that reflects the essential cohesive properties of the source text.

6 Conclusions and Future Work

We have presented an initial collaboration between coherence-based and cohesion-based approaches to discourse processing. We have shown that the information obtained by Lexical Chains and Rhetorical Structure can be successfully combined, yielding a comprehensive representation of discourse. Our initial hypothesis was that this enriched representation of a text should be useful in a knowledge-requiring NLP task such as TS. Indeed, we have shown that it increases compression of the obtained summaries while maintaining precision and recall figures, even when only little information of each system is used by the other.

Analysis of the results reveals that a deeper interaction between these two systems would yield considerable improvements in the final summaries. This includes working with the full version of both systems (see Section 3).

However, we have not been able to adequately assess the improvements achieved by the intensive exploitation of discourse information. One of the main problems is that qualitative aspects of summaries are not adequately captured by the quantitative metrics that we have used in this experiment. Yet, most of the proposed alternative metrics focus in content representativity, and ignore coherence properties of texts. To assess this, task-based evaluations with human judges can be performed. However, drawbacks to experiments with humans is their high cost and polemic evaluation procedures.

Further work includes:

- **Integrating** the observed interactions between lexical chains and rhetorical structure
- **Exploring** further interactions between these two kinds of discursive information
- **Assessing** the improvements that different techniques may introduce in summaries by way of adequate evaluation.

7 Acknowledgements

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

Bibliografía

Alonso, Laura. 2001. A proposal for an automated text summarisation system:integrating rhetorical structure theory and entity detection and tracking. En S. Chaudiron y C. Fluhr, editores, *III Colloque d'ISKO-France: Filtrage et rsum automatique de l'information sur les rseaux.*

Atserias, Jordi, Irene Castellón, y Montse Civit. 1998. Syntactic parsing of unrestricted spanish text. En *First International Conference on Language Resources and Evaluation*, Granada. LREC.

Baldwin, Breck, Robert Donaway, Eduard Hovy, Elizabeth Liddy, Inderjeet Mani, Daniel Marcu, Kathleen McKeown, Vibhu Mittal, Marc Moens, Dragomir Radev, Karen Sparck Jones, Beth Sundheim, Simone Teufel, Ralph Weischedel, y Michael White. 2000. An evaluation road map for summarization research. TIDES, TIDES.

Baldwin, Breck y Thomas S. Morton. 1998. Dynamic coreference-based summarization. En *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing*, Granada, Spain, June.

Banko, Michele, Vibhu Mittal, Mark Kantrowitz, y Jade Goldstein. 1999. Generating extraction-based summaries from hand-written summaries by aligning text spans. En *Proceedings of PACLING-9*, Waterloo, Ontario, July.

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

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

Carbonell, Jaime G. y Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. En *Proceedings of SIGIR*, páginas 335–336.

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

Conroy, John M., Judith D. Schlesinger, Dianne P. O'Leary, y Mary Ellen Okurowski. 2001. Using HMM and Logistic Regression to generate extract summaries for duc. En *Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001*, New Orleans, Louisiana.

Donaway, Robert L., Kevin K. Drummey, y Laura A. Mather. 2000. A comparison of rankings produced by summarization evaluation measures. En *Proceedings of the ANLP/NAACL-2000 Workshop on Automatic Summarization*, páginas 69–78, Seattle, Washington.

Fuentes, Maria y Horacio Rodríguez. 2002. Using cohesive properties of text for automatic summarization. En *JOTRI'02*.

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

Green, Stephen J. 1997. *Automatically generating hypertext by computing semantic similarity.* Ph.D. tesis, University of Toronto.

Grosz, Barbara y Candance Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 3(12):175–204.

Halliday, Michael A. K. y Ruqaiya Hasan. 1976. *Cohesion in English.* English Language Series. Longman Group Ltd.

Hasan, Ruqaiya, 1984. *Understanding Reading Comprehension*, capítulo Coherence and Cohesive Harmony. IRA: Newark, Delaware.

Hearst, Marti. 1994. Multi-paragraph segmentation of expository text. En *32nd Annual Meeting of Association for Computational Linguistics*.

Hirst, Graeme y David St-Onge. 1997. Lexical chains as representation of context for the detection and correction of malapropisms. En Christiane Fellbaum, editor, *WordNet: An electronic lexical database and some of its applications*. The MIT Press, Cambridge, MA.

Hovy, E. y C-Y. Lin. 1999. Automated text summarization in summarist. En Mani y Maybury, editores, *Advances in Automatic Text Summarization*.

Jing, Hongyan. 2001. *Cut-and-Paste Text Summarization*. Ph.D. tesis, Graduate School of Arts and Sciences, Columbia University.

Kan, Min-Yen y Kathleen McKeown. 1999. Information extraction and summarization: Domain independence through focus types. Informe técnico, Computer Science Department, Columbia University, New York.

Knight, Kevin y Daniel Marcu. 2000. Statistics-based summarization - step one: Sentence compression. En *The 17th National Conference of the American Association for Artificial Intelligence AAAI'2000*, Austin, Texas.

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

Kraaij, W., M. Spitters, y M. van der Heijden. 2001. Combining a mixture language model and naive bayes for multi-document summarisation. En *Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001*, New Orleans, Louisiana.

Lin, Chin-Yew y Eduard Hovy. 1997. Identifying topics by position. En *Proceedings of the Applied Natural Language Processing Conference (ANLP-97)*, páginas 283–290, Washington, DC.

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

Marcu, Daniel. 1997a. From discourse structures to text summaries. En Mani y Maybury, editores, *Advances in Automatic Text Summarization*, páginas 82 – 88.

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

McKeown, Kathleen R., Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Barry Schiffman, y Simone Teufel. 2001. Columbia multi-document summarization: Approach and evaluation. En *Proceedings of the Workshop on Text Summarization, ACM SIGIR Conference.*

Muresan, S., E. Tzoukermann, y J. Klavans. 2001. Combining linguistic and machine learning techniques for email summarization. En *ACL-EACL'01 CoNLL Workshop.*

Ono, K., K. Sumita, y S. Miike. 1994. Abstract generation based on rhetorical structure extraction. En *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, páginas 344 – 348, Kyoto, Japan.

Padró, Lluís. 1997. *A Hybrid Environment for Syntax-Semantic Tagging.* Ph.D. tesis, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona.

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

Radev, Dragomir R., Hongyan Jing, y Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. En *ANLP/NAACL Workshop on Summarization*, Seattle, Washington.

Schlesinger, J.D., D.J. Baker, y R.L. Donaway. 2001. Using document features and statistical modeling to improve query-based summarization. En *Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001*, New Orleans, Louisiana.

Sparck-Jones, Karen. 2001. Factorial summary evaluation. En *Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001*, New Orleans, Louisiana.

Tzoukermann, E., S. Muresan, y J. Klavans. 2001. Gist-it: Summarizing email using linguistic knowledge and machine learning. En *ACL-EACL'01 HLT/KM Workshop.*

White, Michael, Tanya Korelsky, Claire Cardie, Vincent Ng, David Pierce, y Kiri Wagstaff. 2001. Multi-document summarization via information extraction. En *Proceedings of the First International Conference on Human Language Technology Research.*