A new hybrid summarizer based on Vector Space model, Statistical Physics and Linguistics

Iria da Cunha[◇], Silvia Fernández[δ,▽], Patricia Velázquez Morales, Jorge Vivaldi[◇], Eric SanJuan[δ] and Juan Manuel Torres-Moreno[∂,δ,⋆⋆]

[◇]Institute for Applied Linguistics, Universitat Pompeu Fabra, Barcelona, España.
{iria.dacunha,jorge.vivaldi}@upf.edu
[δ]Laboratoire Informatique d'Avignon, BP1228, 84911 Avignon Cedex 9, France.
{silvia.fernandez,eric.sanjuan,juan-manuel.torres}@univ-avignon.fr
[∂] École Polytechnique de Montréal/DGI, Montréal (Québec), Canada.
[▽] Laboratoire de Physique des Matériaux, CNRS UMR 7556, Nancy, France.

Abstract. In this article we present a hybrid approach for automatic summarization of Spanish medical texts. There are a lot of systems for automatic summarization using statistics or linguistics, but only a few of them combining both techniques. Our idea is that to reach a good summary we need to use linguistic aspects of texts, but as well we should benefit of the advantages of statistical techniques. We have integrated the Cortex (Vector Space Model) and Enertex (statistical physics) systems coupled with the Yate term extractor, and the Disicosum system (linguistics). We have compared these systems and afterwards we have integrated them in a hybrid approach. Finally, we have applied this hybrid system over a corpora of medical articles and we have evaluated their performances obtaining good results.

1 Introduction

Nowadays automatic summarization is a very prominent research topic. This field has been investigated since the sixties, when techniques based on the frequency of terms [19] or on cue phrases [10] were used. Afterwards other techniques, using textual positions [7, 18], Bayesian models [16], Maximal Marginal Relevance [12] or discourse structure [22, 23, 27] were used. In this work, we focus in medical summarization. We do that because, as [1] indicates, nowadays this is a very important area with a very big amount of information that should be processed, so our work aims to help to solve this problem. As well, we are interested in analyzing the techniques used to summarize texts of specialized areas, specifically the scientific-technical ones, so in the future we will extend this work to other domains as chemistry, biochemistry, physics, biology, genomics, etc. We work with the genre of medical papers because this kind of texts are published in journals with their corresponding abstracts written by the authors, and we employ them to compare with the summaries of our systems in order to carry out the final

⋆⋆ Corresponding author.

evaluation. Another motivation to carry out this work is that, although there are a lot of systems for automatic summarization using statistics [6, 16, 31] or linguistics [2, 22, 27, 33], there are only a few of them combining both criteria [2, 3, 20, 26]. Our idea is that to arrive to a good summary we need to use linguistic aspects of texts, but as well we should benefit of the advantages of statistical techniques. On the basis of this idea, we have developed a hybrid system that takes profit of different aspects of texts in order to arrive at their summaries. To do this, we have integrated three models in this system. Cortex is based in statistics [34], Enertex is based on the Textual Energy [11] and Disicosum is a semiautomatic summarization system that integrates different linguistic aspects of the textual, lexical, discursive, syntactic and communicative structure [9]. In this paper, we have compared these three systems, and afterwards we have integrated them in our hybrid system. Finally, we have applied this system over a corpora of medical articles in Spanish. The resulting summaries have been evaluated with ROUGE [17] obtaining good results. We present a brief experiment in order to observe the influence of the annotator in the tagging process. In Section 2 we describe the three systems that our hybrid system includes. In Section 3 we explain how their integration was carried out. In Section 4 we present the experiments and evaluation, and in Section 5 some conclusions are extracted.

2 Systems used in our Hybrid Approach

2.1 Vector Space Models combining Term Extraction

We tested two different methods for document summarizing, both are based on the Vector Space Model (VSM) [29]. The first method, Cortex, supposes that word frequency can be estimated on the whole set of documents represented as an inverted file.Enertex is inspired in statistical physics, codes a document as a system of spins, and then it computes the "Textual Energy" between sentences to score them. Both systems are coupled with Yate, a term extraction system in order to improve performances in the sentences extraction task.

Cortex system (*Cortex es Otro Resumidor de TEXtos*) [34] is a single-document extract summarization system using an optimal decision algorithm that combines several metrics. These metrics result from processing statistical and informational algorithms on the VSM representation. In order to reduce the complexity, a preprocessing is performed on the topic and the document: words are filtered, lemmatized or/and stemmed. A representation in bag-of-words produces a $S[P \times N]$ matrix of frequencies/absences of $\mu = 1, ..., P$ sentences (rows) and a vocabulary of $i = 1, ..., N$ terms (columns). This representation will be used in Enertex system as well. Cortex system can use up to Γ=11 metrics [34] to evaluate the sentence's relevance. Some metrics are the angle between the title and each sentence, metrics using the Hamming matrix (matrix where each value represents the number of sentences in which exactly one of the terms i or j is present), the sum of Hamming weights of words per segment, the Entropy,

the Frequency, the Interactions and others. The system scores sentences with a decision algorithm combining the normalized metrics. Two averages are calculated, a positive one $\lambda_s > 0.5$ and a negative one $\lambda_s < 0.5$ tendencies ($\lambda_s = 0.5$ is ignored). The decision algorithm combining the vote of each metric is:

$$\sum \alpha = \sum_\nu (||\lambda_s^\nu|| - 0.5); ||\lambda_s^v|| > 0.5; \sum \beta = \sum_\nu (0.5 - ||\lambda_s^\nu||); ||\lambda_s^v|| < 0.5 \quad (1)$$

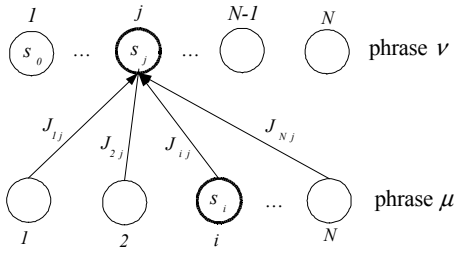
The attributed value to every sentence s is calculated in the following way:

$$\text{if } (\sum \alpha > \sum \beta) \text{ then Score}_s \quad = 0.5 + \frac{\sum \alpha}{\Gamma} : \text{retain } s$$
$$\text{else Score}_s = 0.5 - \frac{\sum \beta}{\Gamma} : \text{not retain } s$$

Γ is the number of metrics and ν is the index of the metrics.

Enertex system [11] is a Neural Network (NN) approach, inspired by statistical physics, to study fundamental problems in Natural Language Processing, like automatic summarization and topic segmentation. The algorithm models documents as a Neural Network whose Textual Energy is studied. The principal idea is that a document can be treated as a set of interacted units (the words) where each unit is affected by the field created by the others. The associative

Fig. 1. Field created by terms of the phrase μ affects the N terms of the phrase ν.



memory of Hopfield [15] is based on physical systems like the magnetic model of Ising (formalism of statistical physics describing a system with two states units named spins) to build a NN able to store/recovery patterns. Learning is following by Hebb's rule [14]:

$$J_{i,j} = \sum_{\text{sentences}} s_i s_j \quad (2)$$

and the recovery by the minimisation of the energy of Ising model [14]:

$$E^{\mu,\nu} = \sum s_i^\mu J_{i,j} s_j^\nu \quad (3)$$

The main limitation of Hopfield NN is its storage capacity: patterns must be not-correlated to obtain free error recovery. This situation strongly restricts its

applications, however Enertex exploits this behaviour. VSM represents document sentences into vectors. These vectors can be studied as a NN. Sentences are represented as a chain of N active (present term) or inactive (absent term) neurons with a vocabulary of N terms per document (Fig. 1). A document of P sentences is formed by P chains in a N dimension vector space. These vectors are correlated, according to the shared words. If topics are close, it is reasonable to suppose that the degree of correlation will be high. We compute the interaction between terms by using (2) and the Textual Energy between phrases by (3). Weights of sentences are obtained using their absolute values of energy. The summary consists of the relevant sentences having the biggest values.

Yate system The terms extracted by this tool represent "concepts" belonging to the domain found in the text and their termhood will modify the weights in the term-segment matrix. Yate [35] is a term candidate extraction tool whose main characteristics are: a) it uses a combination of several term extraction techniques and b) it uses EWN[1], a general purpose lexico-semantic ontology as a primary resource. Yate was designed to obtain all the terms (from the following set of syntactically filtered candidates: <noun>, <noun-adjective> and <noun-preposition-noun>) found in Spanish specialised texts within the medical domain. Yate (Fig. 2) is a hybrid tool that combines the results obtained by a set of term candidate analysers: a) domain coefficient: it uses the EWN ontology to sort the term candidates[2], b) context: it evaluates each candidate using other candidates present in its sentence context, c) classic forms: it tries to decompose the lexical units in their formants, taking into account the formal characteristics of many terms in the domain and d) collocational method: it evaluates multiword candidates according to their mutual information. The results obtained by this set of heterogeneous methods are combined to obtain a single list of sorted term candidates [35].

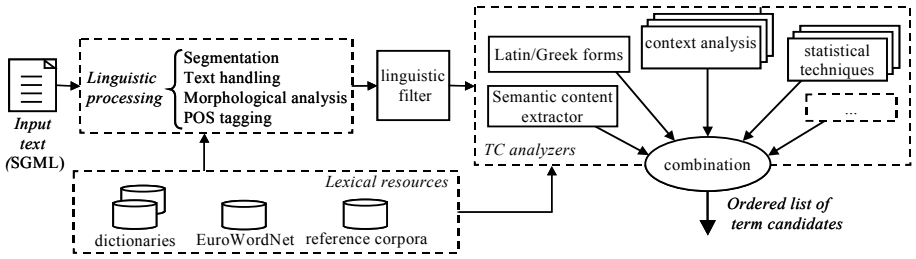
2.2 Linguistic model: Disicosum system

The conception of this summarization model of medical articles was done under the hypothesis that professionals of specialized domains (specifically, the medical domain) employ concrete techniques to summarize their texts [9]. [9] have studied a corpora containing medical articles and their abstracts in order to find

[1] EWN (www.illc.uva.nl/EuroWordNet) is a multilingual extension of WordNet (wordnet.princeton.edu), a lexico-semantic ontology. The basic semantic unit in both resources is the "synset" that groups together several single/multi words that can be considered synonyms in some contexts. Synsets are linked by means of semantic labels. Due to polysemy, a single lexical entry can be attached to several synsets.

[2] This module locates zones in EWN with high density of terms. The precision obtained when performing in isolation depends on many factors, such as the degree of polysemy of the term candidate or the relative density of terms of the EWN zone. The coverage is high with the obvious limitation of being all or part of the components of the term candidate in EWN. See [36] for details.

Fig. 2. Architecture of Yate.



which kind of information should be selected for a specialized summary and, afterwards, to do generalizations to be included in their model of summarization. Another starting point of this model was the idea of that different types of linguistic criteria should be used to have a good representation of texts, and in this way exploit the advantages of each criteria. This idea is quite new because, generally, automatic summarization systems based on linguistics use one type of criteria (as we have mentioned above, terms in [19]; textual position in [7, 18]; discursive structure in [22, 33], etc.), but not the combination of different linguistic criteria. [9]have found linguistic clues that come from the textual, lexical, discursive, syntactic and communicative structures. The system is formed by rules concerning each of those five structures. In the first place, the textual rules of the system indicate that: i) The summary should contain information from each section of the article: Introduction, Patients and methods, Results and Conclusions [32]. ii) Sentences in the following positions should be given an extra weight: the 3 first sentences of the Introduction section, the 2 first sentences of the Patients and methods and the Results sections, and the 3 first and the 3 last sentences of the Conclusions section. In the second place, the system contains lexical rules of two types: a) Lexical rules increasing the score to sentences containing: i/ Words of the main title (except stop words), ii/ Verbal forms in 1st plural person, iii/ Words of a list containing verbs (*to analyse*, *to observe*, etc.) and nouns (*aim*, *objective*, *summary*, *conclusion*, etc.) that could be relevant, iv) Any numerical information in the Patients and method and the Results sections. b) Lexical rules eliminating sentences containing: i/ References to tables/figures (linguistic patterns show that only a part of sentence should be eliminated: *As it is shown in Table... In Figure 4 we can observe...*), ii/ References to statistical/computational aspects: *computational*, *program*, *algorithm*, *coefficient*, etc., iii/ References to previous work: *et al* and some linguistic patterns, for example, "determinant + noun (work|study|research|author)". Exceptions: *this study*, *our research...* iv/ References to definitions: *it is/they are defined by/as....*

Finally, the system includes discursive rules and rules combining discursive structure with syntactic and communicative structure. In order to formalize these rules we follow two theoretical frames: the Rethorical Structure Theory (RST) [21] and the Meaning-Text Theory (MTT) [24, 25]. The RST is a theory of the

organization of the text that characterizes its structure as a hierarchical tree containing discursive relations (Elaboration, Concession, Condition, Contrast, Union, Evidence, etc.) between its different elements, that are called nucleus and satellites. The MTT is a theory that integrates several aspects of the language. In our work, on the one hand, we use its conception of the deep syntax of dependencies, that represents a sentence as a tree where lexical units are the nodes and the relations between them are marked as Actants and the Attributive, Appenditive and Coordinative relations. On the other hand, we use the distinction between Theme and Rheme, that is part of the communicative structure of the MTT. Some examples of the these rules are[3]:

- IF S is satellite$_{CONDITION}$ C THEN KEEP S
  [If these patients require a superior flow,] S [it is probably that it is not well tolerated.] N
- IF S is satellite$_{BACKGROUND}$ B THEN ELIMINATE S
  ~~[Persons who don't want eat and with a complex of fatness have anorexia.]~~ S [We have studied the appearance of complications in anorexic patients.] N
- IF S is satellite$_{ELABORATION}$ E1 AND S elaborates on the Theme of the nucleus of E1 THEN ELIMINATE S
  [Persons who don't want eat and with a complex of fatness have anorexia.] N ~~[One of the problems of these patients is the lack of self-esteem.]~~ S
- IF S is satellite$_{ELABORATION}$ E1 AND S is ATTR THEN ELIMINATE S
  [They selected 274 controls,] N ~~[that hypothetically would have had the same risk factors.]~~ S

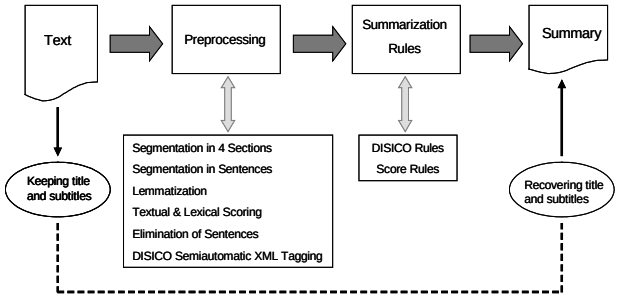
2.3 Limitations and Solutions for the Model's Implementation

For the implementation of the textual and lexical rules of the model there are no problems because there are preanalysis tools: a segmentation tool developed at the Institute for Applied Linguistics and the Spanish TreeTagger [30]. But we found some problems for the full implementation of the model. The first one, is that there are no parsers able to obtain the discursive structure in Spanish texts. There is one [22, 23] for English, and a current project for the Portuguese [28]. The second one, is that there is not known parser to obtain the communicative structure in any language. There are only a few publications about it, as for example [13]. The third one, is that, although there are some syntactic parsers of dependencies for Spanish [5, 4], their results, at the moment, are not so reliable as the system needs. So, the solution was to simulate the output of these parsers. Thus, a semiautomatic discursive and syntactic-communicative XML tagger was designed, in order to tag texts and afterwards apply the linguistic summarization model over them [8]. To tag these texts, an annotation interface, where the user can choose the relation between the different elements (nucleus and satellites) of the texts, has been developed. It has to be taken into account that this tagging can be done in two stages. First, the user can detect the relations between sentences and, afterwards, he may find more relations inside each sentence (if any). The final result will be a representation of the text in form of a relations

[3] N=nucleus, S=satellite. Text underlined will be eliminated.

tree, over which the summarization system will be applied. Figure 3 shows the architecture of Disicosum. It is necessary to point out that the rules of the model are applied over the text of each section separately.

Fig. 3. Architecture of the medical summarization model.



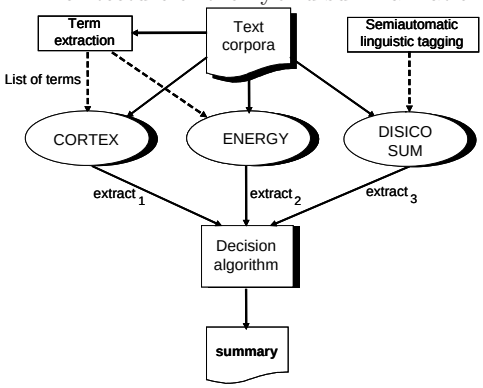
3 Our Hybrid Approach of Automatic Text Summarization

We have already presented the main characteristics of the different components that our hybrid system integrates. This section briefly will show how such components are integrated in a single hybrid summarization system. Figure 4 presents the system architecture. Firstly, the system applies the elimination rules presented in Section 2.2 over the original text, which produces a reduction $\approx$ 20-30% in its length. Over this reduced text, the system applies separately the Cortex, Enertex and Disicosum systems. A Decision Algorithm processes the normalized output of systems as follow: in the first step, the algorithm chooses the sentences selected by the three systems. If there are no consensus, it chooses the sentences selected by two systems. Finally, if there are sentences selected by only one system, the algorithm gives priority to the sentences with the biggest score.

4 Experiments and evaluation

The corpora used for testing contains 10 Spanish medical papers of the *Medicina Clínica* journal. Before applying the system, texts were semiautomatically tagged with the developed interface by 5 people (2 texts each one). Afterwards, we have compared the summaries produced by the 3 systems, giving to them as input the articles reduced by applying the elimination rules (c.f. 2.2). Also we have created 2 baselines in order to include them in the performance comparison. The difference between them was that Baseline_1 was made from the original article, and Baseline_2 was made from the original article reduced by the application of

Fig. 4. Architecture of the hybrid summarization system.



the elimination rules mentioned above (2.2). To evaluate the summaries we have compared them with the abstracts written by the authors, using ROUGE [17]. In order to interpret the results, it has to be taken into account that authors' summaries are abstracts, while the summaries of our system are extracts. For the application of ROUGE, we have used a Spanish lemmatization and a stop-word list. To set the length of the summaries, we have computed the average number of sentences in each section, present in the author's summaries. Then, we decided to include in our summaries one additional sentence per section. This decision was made because we have noticed that usually authors give, for example, one sentence with different contents in their abstracts, but in their articles they give those contents in separate sentences. In short, it was an empirical decision in order to not lose information. Finally, the system chooses 2 sentences for Introduction and Discussion sections, 3 sentences for the Patients and methods section, and 4 sentences for Results section (11 sentences altogether). In order to analyze the performance of the hybrid system that we present in this article, we have applied it over the ten articles of our corpora, obtaining their summaries. The evaluation results are shown in Table 1. We have used ROUGE measures despite the fact that only one reference abstract is provided. Nevertheless, ROUGE measures provide a standard way of comparing abstracts based on bi-grams and guarantees the reproducibility of our experiments. To evaluate the impact of the quality of semi-automatic tagging on Disicosum performance, two documents among the ten were tagged in a restricted time (30 min per text) and the others without time restrictions. Therefore, the coherence of the linguistic tagging on these texts is expected to be better than for the two texts tagged in restricted time. Table 1 gives the median score of each system on the ten documents (column median1) and on the reduced set of documents tagged without time restrictions (column median2). Font scores depend on the quartile (big fonts for higher quartiles, smaller ones for others). Regarding the median score on the documents which tagging has been done in an unrestricted time, Disicosum abstracts seem to be the closest to author abstract according to ROUGE-2 measure.

Table 1. Comparison the ROUGE values between different summaries.

	ROUGE-2		SU4	
System	**Median 1**	**Median 2**	**Median 1**	**Median 2**
Hybrid system	**0.3638**	**0.3808**	**0.3613**	**0.3732**
Disicosum	**0.3572**	**0.3956**	**0.3359**	**0.3423**
Cortex	**0.3324**	0.3324	**0.3307**	**0.3255**
Enertex	0.3105	0.3258	0.3155	0.3225
Cortex on full text	0.3218	**0.3329**	0.3169	0.3241
Enertex on full text	**0.3598**	**0.3598**	**0.3457**	**0.3302**
Baseline_1	0.2539	0.2688	0.2489	0.2489
Baseline_2	0.2813	0.3246	0.2718	0.3034

According to SU-4 Disicosum is the best among the individual systems but the hybrid system has a better score. Regarding the whole set of texts, the median score of Disicosum is lower than the previous one, however it remains among the higher ones for individual systems. This shows that the quality of the linguistic model tagging has a direct impact on the summary quality meanwhile the tagging is carried out independently from the summarisation purpose. Cortex and Enertex systems have been tested directly on full texts or after segmenting texts into independent sections. The segmentation preprocess is part of Disicosum. The second best individual system according to these results seems to be Enertex on full text. It appears that Enertex works better without the indication that the summary should contain elements coming from each section of the text. An explanation could be that Enertex compares all sentences two by two. The more sentences the text has, the better is the vector representation of the sentence in the system. Cortex uses more local criteria since it has been built to efficiently summarise large corpora. On short texts, the lack of frequent words reduces the efficiency of the system however it appears here that it can take into account the structural properties of the texts. Looking at the hybrid system, the experiment shows that it improves the proximity with the author's abstract in all cases except for ROUGE-2 when considering human linguistic tagging done without time restriction. Finally, we have carried out another experiment: we use the same text (number 6) annoted by five different people and theirs summaries generated by Disicosum as reference models, and we have decided to compute ROUGE tests over all other systems. The idea is to find which system is closer to models. Results and an example of summary are shown in Table 2. Cortex and Enertex are the closer systems to the linguistic model. In other words, performance of Disicosum and the two numerical summarizers used are equivalents.

5 Conclusions

We show in this paper, on the one hand, that the summaries produced by statistical methods (Cortex and Enertex) are similar to the summaries produced by

Table 2. ROUGE values for five different summaries models and example of summary.

	ROUGE-2	SU4
Author	0.1415	0.1710
Cortex	**0,7281**	**0.7038**
Enertex	**0,7281**	**0.7038**
$Baseline_1$	0.3059	0.2920
$Baseline_2$	0.3662	0.3740

Evaluación de las vías de acceso venoso innecesarias en un servicio de urgencias. Fundamento. *Los accesos venosos son uno de los proce dimientos que con más frecuencia se practican en los servicios de urgencias con un fin terapéutico, que puede ser inmediato o no, en función de la sintomatología que presente el paciente o el diagnóstico de sospecha inicial. El objetivo del presente trabajo fue evaluar el volumen de pacientes a quienes se les practica un acceso venoso, estimar cuántos de ellos son innecesarios y el coste económico que ello genera.*

linguistic methods. On the other hand, we have proved that combining statistics and linguistics in order to develop a hybrid system for automatic summarization gives good results, even better than those ones obtained by each method (statistical or linguistic) separately. Finally, we have tested that the Disicosum system offer very similar summaries although different annotators tag the original text (that is, the annotators give different discourse trees). Other tests, comparing several summaries produced by doctors and our hybrid system, may be realized. Extensions to other domains and languages will also be considered.

References

1. S. Afantenos, V. Karkaletsis, and P. Stamatopoulos. Summarization of medical documents: A survey. *Artificial Intelligence in Medicine*, 2(33):157–177, 2005.
2. L. Alonso and M. Fuentes. Integrating cohesion and coherence for Automatic Summarization. In *EACL'03 Student Session*, pages 1–8. ACL, Budapest, 2003.
3. M. Aretoulaki. *COSY-MATS: A Hybrid Connectionist-Symbolic Approach To The Pragmatic Analysis Of Texts For Their Automatic Smmarization.* PhD thesis, University of Manchester, Institute of Science and Technology, Manchester, 1996.
4. J. Asterias, E. Comelles, and A. Mayor. TXALA un analizador libre de dependencias para el castellano. *Procesamiento del Lenguaje Natural*, 35:455–456, 2005.
5. G. Attardi. Experiments with a Multilanguage Non-Projective Dependency Parser. In *Tenth Conference on Natural Language Learning*. New York, 2006.
6. R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Intelligent Scalable Text Summarization Workshop, ACL, Madrid, Spain.*, 1997.
7. R. Brandow, K. Mitze, and L. Rau. Automatic condensation of electronic publications by sentence selection. *Inf. Proc. and Management*, 31:675–685, 1995.
8. I. da Cunha, G. Ferraro, and T. Cabre. Propuesta de etiquetaje discursivo y sintáctico-comunicativo orientado a la evaluación de un modelo lingüístico de resumen automático. In *Conf. Asoc. Española de Lingüística Aplicada*. Murcia, 2007.
9. I. da Cunha and L. Wanner. Towards the Automatic Summarization of Medical Articles in Spanish: Integration of textual, lexical, discursive and syntactic criteria. In *Crossing Barriers in Text Summarization Research*. RANLP, Borovets, 2005.
10. H. P. Edmundson. New Methods in Automatic Extraction. *Journal of the Association for Computing Machinery*, 16:264–285, 1969.
11. S. Fernández, E. SanJuan, and J. M. Torres-Moreno. Énergie textuelle de mémoires associatives. *Traitement Automatique des Langues Naturelles*, pages 25–34, 2007.
12. J. Goldstein, J. Carbonell, M. Kantrowitz, and V. Mittal. Summarizing text documents: sentence selection and evaluation metrics. In *22nd Int. ACM SIGIR Research and development in information retrieval*, pages 121–128. Berkeley, 1999.

13. E. Hajicova, H. Skoumalova, and P. Sgall. An Automatic Procedure for Topic-Focus Identification. *Computational Linguistics*, 21(1), 1995.
14. J. Hertz, A. Krogh, and G. Palmer. *Introduction to the theorie of Neural Computation*. Redwood City, CA : Addison-Wesley, 1991.
15. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *National Academy of Sciences*, 9:2554–2558, 1982.
16. J. Kupiec, J. O. Pedersen, and F. Chen. A trainable document summarizer. In *SIGIR-95*, pages 68–73. New York, 1995.
17. C. Lin. Rouge: A Package for Automatic Evaluation of Summaries. In *Workshop on Text Summarization Branches Out (WAS 2004)*, pages 25–26, 2004.
18. C. Lin and E. Hovy. Identifying Topics by Position. In *ACL Applied Natural Language Processing Conference*, pages 283–290. Washington, 1997.
19. H. P. Luhn. The automatic creation of Literature abstracts. *IBM Journal of research and development*, 2(2), 1959.
20. A. M. Towards a Hybrid Abstract Generation System. In *Int. Conf. on New Methods in Language Processing*, pages 220–227. Manchester, 1994.
21. W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.
22. D. Marcu. *The rhetorical parsing, summarization, and generation of natural language texts*. PhD thesis, Dep. of Computer Science, University of Toronto, 1998.
23. D. Marcu. *The Theory and Practice of Discourse Parsing Summarization*. Institute of Technology, Massachusetts, 2000.
24. I. Mel'cuk. *Dependency Syntax: Theory and Practice*. Albany: State University Press of New York, 1988.
25. I. Mel'cuk. *Communicative Organization in Natural Language. The semantic-communicative structure of sentences*. John Benjamins, Amsterdam, 2001.
26. T. Nomoto and Y. Nitta. A Grammatico-Statistical Approach to Discourse Partitioning. In *15th Int. Conf. on Comp. Linguistics*, pages 1145–1150. Kyoto, 1994.
27. K. Ono, K. Sumita, and S. Miike. Abstract generation based on rhetorical structure extraction. In *15th Int. Conf. on Comp. Linguistics*, pages 344–348. Kyoto, 1994.
28. T. Pardo, M. Nunes, and M. Rino. DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. In *SBIA2004*, pages 224–234. São Luís, 2004.
29. G. Salton and M. McGill. *Introduction to modern information retrieval*. Computer Science Series McGraw Hill Publishing Company, 1983.
30. H. Schmid. Probabilistic Part-of-speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, 1994.
31. H. G. Silber and K. F. McCoy. Efficient text summarization using lexical chains. In *Intelligent User Interfaces*, pages 252–255, 2000.
32. J. Swales. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, Cambridge, 1990.
33. S. Teufel and M. Moens. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28, 2002.
34. J. M. Torres-Moreno, P. Velázquez-Morales, and J. G. Meunier. Condensés de textes par des méthodes numériques. In *JADT*, pages 723–734. St. Malo, 2002.
35. J. Vivaldi. *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. PhD thesis, Universitat Politècnica de Catalunya, Barcelona, 2001.
36. J. Vivaldi and H. Rodríguez. Medical term extraction using the EWN ontology. In *Terminology and Knowledge Engineering*, pages 137–142. Nancy, 2002.