Automatic Generation of Document Summaries in Spanish Language

Rodolfo Rodríguez, Darnes Vilariño, Beatriz Beltrán, and Mireya Tovar

Benemérita Universidad Autónoma de Puebla, Puebla, Puebla, México.
Facultad Ciencias de la Computación
Puebla, Puebla, 72570 México, www.cs.buap.mx

Abstract. Without a doubt, Internet has become the biggest source of available information. Nowadays one of the biggest sources of information is the WEB, the information grows in a chaotic and not controlled way, originating certain limitations for its handling, organization and recovery. The Spanish language is very complicated for its study, and exist few tools that make an analysis to obtain an abstract and so the few available versions unfortunately are not freeware. In the present work we developed a learning technique and a set of metrics that applied at the original document return us the most representative sentences of the document to construct the automatic abstract by extraction, including a study of anaphoras.

Keywords. Retrieval, information, abstract, automatic, extraction, sentence, learning.

1 Introduction

A summary is to reduce to brief terms the essential of a document. The use of the summary of a document helps to reduce the storage space, facilitates the access to the most important information and accelerates the time of reading to locate the required information of a particular document, in at a certain moment [11].

Nowadays it is necessary to count on suitable computer science tools for the recovery of the important information in an efficient and fast way. The manual techniques have demonstrated to be inefficient, because basically they consist in manual elaboration. This work is very expensive in time, in addition it suffers from abundant inconsistencies, like orthography and coherence [3,4]. Now identified the problem, the challenge of the investigators is achieve that the computers be an efficient tool in the process of information retrieval.

By automatic generation of text summaries is understood the process by which the originating substantial information of a source (or several) is identified to produce a brief version destined to a particular user (or user group) and a task (or tasks) [22].

A Phrase and an sentence are different in which the first lacks of a express preaching (does not say anything of a subject, type: Buenos Días) and in the second

© *A. Gelbukh, M. Adiba. (Eds.)*
Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 185-197

Received 31/05/08
Accepted 04/08/08
Final version 10/09/08

there is explicit preaching (one affirms or it denies, something is said from somebody or something: Buenos Días nos de Dios (Subject: Dios; Predicate: de).

Sciences and the philosophy tend to establish truth, but the concept of true - or false - is not possible apply it to smaller units of the language than the sentences: nor to the sound. Therefore the minimum unit of the susceptible language of true is the proposal, since all proposal or sentence implies a judgment (establishment of the truth or false principle of the knowledge) [21].

There are a lot of articles that propose methods to extract the most important orations of a determined document, using the frequency of the words within the same document, the frequency of words within the title, considering most important the first five and last five sentences of a document, as well as other statistical parameters [4,7]. Nevertheless the great majority of these articles does not work over the Spanish language, the most important results are on the English language, that is substantially different with the Spanish language.

In the present article we obtained automatic summaries by extraction in the Spanish language are, where some classic techniques like eliminate stopwords and special symbols are used, stemm the words, make a general study of the most well-known abbreviations, as well as a brief study of proper names in a pre processed, also we used some other metrics for the selection of the important sentences, as they are: eliminate very short sentences, important words, comparison with the title, etc., in addition we developed a technique that includes intelligence (learning) for the important information retrieval of each document and when the extract is obtained, a study of the most well know anaphora [1]is applied with the intention of have more coherency in the obtained extract. All used software has been made by our group of investigation and the techniques used for the English language have been adapted to work in the Spanish language.

The present article is divided in different sections. In the second section, we describe a general way all the different metrics that were used to evaluate each sentence in a document. In the third section, we present the used algorithm to obtain the automatic summary by extraction, where the system decides which metrics will be used to evaluate the document and obtain the summary. The fourth section is dedicated to explain how works the algorithm that include anaphoras. The fifht section of the article is all about to evaluate the system, where the obtained results are show in the first table. This table shows the similarity degree between the obtained summary by the system without anaphoras against the obtained summary for one of the three experts, and also shows the degree of similarity between the summary of the system without anaphoras and with anaphoras, and in the second table we present the obtained results of the threshold for each metric (that help us to decide which metric will be used or not to evaluate a document) after analyzed all the documents of our corpus. The last section are the conlusions.

[1] Anaphora. Word that is used to reference other words that were used before in the document.

2 Score of Sentences of the Document

Supported in the analysis made at the different reviewed articles, it had been collected and adapted to our problem a set of metric that they search to give certain score at each sentence of the document, and extract the five most important sentences (the five sentences with bigger score) to conform an automatic summary by extraction.

2.1 Selection of Important Sentences using the Transition Point (TP)

It is defined the TP like the frequency of a term of the text that divides in two the terms of a vocabulary (in terms of high and low frequency), that is to say, the terms nearest the TP as much of high and low frequency are going to determine of what it is all about the document [12].

From the transition point a determined threshold can be taken, works generally with 25% or with 40%, in this work, words above 25% of TP and under 25% the TP of the threshold are selected, with these terms is constructed a paragraph, it is denominated Virtual Paragraph (VP).

With the aid of the VP, it is given a score at each sentence according to its degree of similarity, applying the similarity formula of Jaccard (1) at each sentence (O_i) of the document, and the obtained VP.

$$Sim_i(PV, O_i) = \left|\frac{PV \cap O_i}{PV \cup O_i}\right| \quad (1)$$

2.2 Sentences Length

This metric is sustained in the idea that the sentences of very small length are not important, because generally they are not in the summaries generated by the humans, by this, it is assign a negative score to them that affects to be part of the automatic summary [10]. In the present work it is considered to be a small sentence, if it consists of less than 5 words, after doing corresponding pre processing.

2.3 Title Compare

The title of a document usually is strongly related to its content and often constitutes the best summary of itself [11].

It is constructed a VP in the same way that was in the TP, but in this case this VP is formed by the terms that conform the title of the document [15]. The VP created from the title words is compare with each one of the sentences of the document, applying the similarity of Jaccard (1), then it is give a weight to each sentence of the

original document, and those with greater weight are candidates to be part of the automatic summary by extraction.

2.4 Centroid

The centroid value is a measurement that indicates how frequent are the words of a document [10, 11, 13, 15]. Importance is attributed to this metric, because using it, is possible know how near are the sentences of a document to the main subject. The centroid is a very great vector; each component of the vector is one of the words that appear in the document. Be D the document and $|D|$ the number of sentences in the document, it is described by:

$$Centroide(D) = (v_{w0}, v_{w1}, \dots, v_{wn}) \tag{2}$$

Where:

$$v_{wi} = \frac{TF(w_i) * IDF(w_i)}{|D|} \tag{3}$$

$TF(w_i)$ Is define like the occurrence number of w_i in the document D.

$IDF(w_i)$ Is known like inverse frequency, and this is calculated by the following formula:

$$IDF(w_i) = \log\left(\frac{|D|}{Sentences\ in\ D\ that\ include\ w_i}\right) \tag{4}$$

IDF indicates how "weird" is the word w_i in the document D.

2.5 Position

This metric is considered important due to the fact that in different types of documents, the most important information occurs in the first sentences, is to say the main idea occurs in the first paragraphs and in the rest of the document this idea simply is developed [10, 11, 13, 15]. For each sentence O_i in the document D, the value of the position of each sentence it is calculate by:

$$P_i = \left(\frac{(|D| - i + 1)}{|D|}\right) * C_{max} \tag{5}$$

Where C_{max} is the maximum centroid value obtained by the previous formula (2).

This metric is combined with the previous one, thus finally the sentence O_t obtain a higher score in function of the position and the obtained centroid value, the sentences importance diminishes with the position.

In addition to the previous score, a greater weight is offers to the first 5 and last 5 sentences of the document.

2.6 Proper Names

For some authors the presence of proper names is important [11, 13, 15]. Is considered that the sentences that have proper names are more important, since it can refer a particular person or place, and this is very common in bibliographies. In this work it has been considered like proper name those words that begin with a capital letter and which are not stopwords.

2.7 Important Words

The important Words metric [10, 11, 14, 17, 18] are words and expressions that do not have relation with the central subject of the document, but they indicate that the sentence can contain important information and must be part of the summary, words like: "importante", "esencial", "en conclusion", "resumen", "fundamental", etc. [11]. As we would desire the most possible generic system, a set of certain words is used, that reveal importance independently of the type of document.

The implemented metric looks for those important words within the text to summarize, granting additional scores to the sentences that contain them.

2.8 Sentences Compare

With this metric we try to find the sentences with greater degree of similarity. Each sentence of the document is compared with all the rest, applying the similarity of Jaccard (1); to define that two sentences are similar, is taken a threshold of 0.5, this is a good measurement (if it be 1 would be the same statement the one that is being compared) for the comparison of two sentences, if the similarity is 0 means that no one of the words that contains the sentence are similar to which is compared. The sentences with a greater degree of similarity are given a bigger score to conform the automatic summary by extraction.

2.9 Bayes Probability

For each sentence s we are going to calculate the probability if it will be included in a summary S given k characteristic, $F_j; j = 1, \ldots, k$, that can be expressed using the Bayes Rule and assuming statistical independence of the metrics, we have:

$$P(s \in S|F_1, F_2, \ldots, F_K) = \frac{\prod_{j=1}^{k} P(F_j|s \in S)P(s \in S)}{\prod_{j=1}^{k} P(F_j)} \quad (6)$$

Where:

- $P(s \in S|F_1, F_2, \ldots, F_K)$ It is the probability that the sentence s is contained in the summary given by the metrics $F_j; j = 1, \ldots, k$.
- $P(s \in S)$ It is the probability that the sentence s is selected between all the others of the document.
- $P(F_1, F_2, \ldots, F_K|s \in S)$ It is a probability that can be considered counting the number of metrics that were used of the total in the sentence s.
- $P(F_j)$ This probability is calculated of the average of the number of times that is used a metric in the document (just it is taken once like maximum account by sentence).

3 Algorithm to Extract Important Sentences

We made and developed an algorithm in order to decide if the j metric will be used or it does not to give an additional score to a sentence s, taking the idea from the Bayes probability, so $P(F_j)$ must be greater to a given threshold to be used; This threshold will be calculated applying the following algorithm.:

1. Fort the first analyzed document we are going to consider that there is the same probability that a metric j is used or not, reason why the initial threshold will be equal to 0.5, i.e. $Threshold_1 = 0.5$.
2. The threshold is calculated for the following documents in this way:

$$Threshold_j = \frac{Threshold_1 + \sum_{i=1}^{k} P_i(F_j)}{k+1} \quad (7)$$

Where k is the number of analyzed documents of the experiment, and i is the number of the analyzed document $i: 1 \ldots k$.

When concluding the experiment of analyzed documents, the threshold will be fixed in the obtained value. In this way if the average of times that was used a metric in a document is smaller than the threshold obtained from that metric, then it will not be used to evaluate the sentences, so the metric will not influence in the decision of the sentences to be included or not in the summary given by the system.

Once all the metrics finished giving the corresponding scores to each one of the sentences, the scores are added and mediated by sentences. The 5 sentences with greater average score are selected and ordered by appearance to construct the automatic summary by extraction.

4 Algorithm to Obtain Anaphoras

Once we have the automatic summary by extraction, we made a study of anaphoras, the main objective is give coherence to the extract made by the system, we did this in the following way:

1. Search for anaphoras in the sentences of the extract obtained by the system.
2. If the system finds a sentence that contains an anaphora, the previous sentence of which was the anaphora will be included in the new extract, this new sentence is taken from the original document.
    a. If the new sentence is already in the original extract do nothing.
3. Include in the new extract the sentence in which was found the anaphora.
4. Do the same for each one of the sentences of the original extract obtained by the system.

There are not methods of resolution of anaphora for the Spanish language that not use much semantic information. In this article one of our main objectives were implement a tool for the resolution of anaphoras for the Spanish language using limited knowledge, it means, not using semantic information.

5 Evaluation

Until now to analyze the behavior of the method of obtaining automatic summaries by extraction, was decided work with a set of 100 heterogeneous documents in the Spanish language, each one with its respective summary of one of a set of three experts.

An automatic pre processing was made at each document, first the stopwords were eliminated, then the document was stemmed and a treatment of abbreviations was made (that consists of changing the most common abbreviations by the complete word), to conform the initial groups.

Between the 100 documents that conform our CORPUS, there are very small documents and greater others, to all documents the formula of similarity of Jaccard (1) was applied to them between the summary offered by the Expert, against the summary obtained by the system, and the summary obtained by the system without anaphoras and with anaphoras, with the objective to measure the proportion of similar sentences, the results are presented in Table 1.

In the Table 1 can be appreciate that the similarity degree given by the extract of the system without anaphoras and the extract by one of the three experts is approximately of a 60%, and the similarity degree between the extract of the system without anaphoras and with anaphoras is approximately of a 98%, this means that many of the sentences of the extract of the system without anaphoras did not have anaphoras or its anaphora was included already in the initial extract, reason why when the additional sentences were included in the new extract, the similarity degree decrement between the extracts.

Table 1. Obtained similarity results

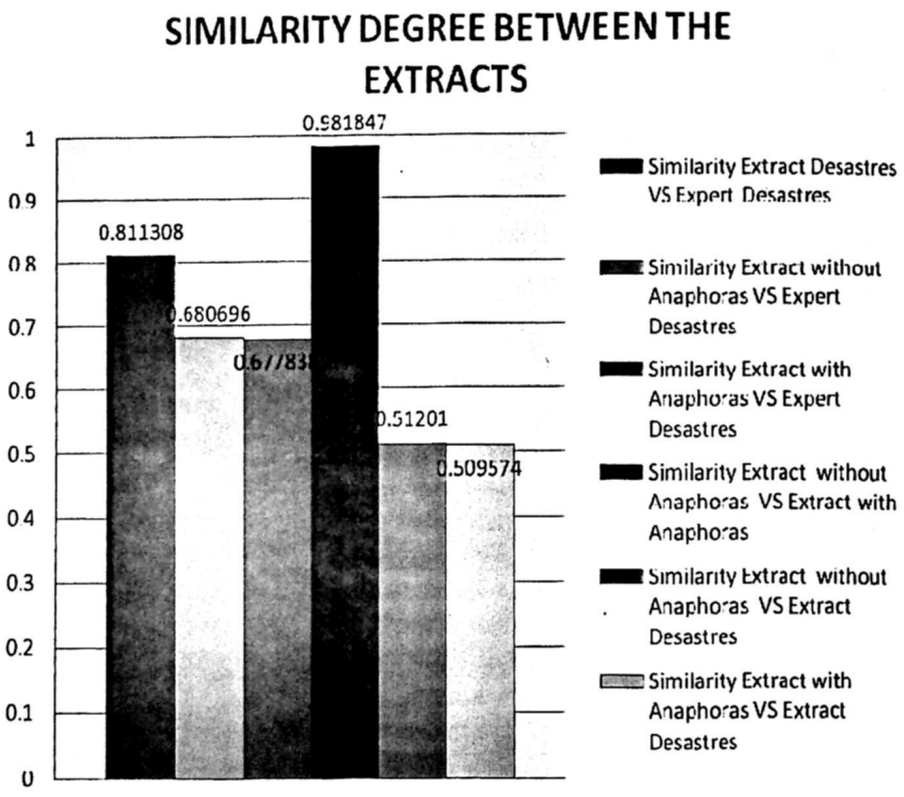
Document	Similarity Extract without Anaphoras VS Expert	Similarity Extract without Anaphoras VS Extract with Anaphoras
n1	0.534031	0.921052
n2	0.735135	1.0
n3	0.629107	1.0
n4	0.663414	1.0
...	...	...
n21	0.40566	0.916334
n22	0.62	1.0
n23	0.769230	0.917391
n24	0.721649	1.0
n25	0.591928	1.0
Antrax	0.593023	1.0
Arqueología	0.690476	1.0
Ciencia	1	1.0
Comercio electrónico	0.5	0.962358
Cristianismo	0.349999	1.0
...	...	...
Genoma humano	0.473282	0.960548
Guanajuato electrónico	0.661710	0.955453
Alimentación	0.581818	0.965442
General Average	**0.599416**	**0.981847**

Table 2. Similarity degree between the extracts

Document Corpus Desastres	Similarity Extract Desastres VS Expert Desastres	Similarity Extract without Anaphoras VS Expert Desastres	Similarity Extract with Anaphoras VS Expert Desastres	Similarity Extract without Anaphoras VS Extract with Anaphoras	Similarity Extract without Anaphoras VS Extract Desastres	Similarity Extract with Anaphoras VS Extract Desastres
Afecta Incendio	1.0	0.437956	0.437956	1.0	0.372262	0.372262
Alarma de Incendios	0.705882	0.462686	0.462686	1.0	0.283105	0.283105
Alerta por un incendio	1.0	0.550387	0.479729	0.907767	0.488372	0.425675
Alerta Roja	0.888482	0.789189	0.789189	1.0	0.607594	0.607594
Amaga incendio	0.607028	0.778688	0.778688	1.0	0.615803	0.615803
Amenazan a Arizona	0.927756	1.0	1.0	1.0	0.784482	0.784482
Amenazan Sydney	0.771929	0.833333	0.778098	0.934844	0.384615	0.350194
Arraso incendio	0.638888	0.769230	0.769230	1.0	0.741258	0.741258
Aumentan daños	0.857142	0.729805	0.729805	1.0	0.672268	0.672268
Ya son 205	0.478260	0.719008	0.719008	1.0	0.0625	0.0625
Visitara Fox zona	1.0	0.929752	0.929752	1.0	0.743801	0.743801
Tormenta Tropical	0.757201	0.585365	0.585365	1.0	0.525252	0.525252
Toneladas de ayuda	0.894977	0.633451	0.633451	1.0	0.528	0.528
Tifon Siembra	0.638115	0693121	0.693121	1.0	0.833638	0.833638
TERREMOTO	0.808510	0.714285	0.714285	1.0	0.561797	0.561797
Temporada de incendios	0.876190	0.946808	0.942105	0.987013	0.655172	0.653409
...	...	...	...	...	...	...
General Average	**0.811308**	**0.680696**	**0.677838**	**0.981847**	**0.512010**	**0.509574**

In Table 2 we made the comparison between the results obtained by our system against the results obtained by the system realized in the master thesis "Automatic Generation of Multiple Document Summaries" [24]. The Corpus with which was made the comparison called DISASTERS was realized in the INAOE (National Institute of Astrophysics Optical and Electronic) and is a corpus specialized in Natural Disasters of news of Mexican newspaper of national circulation; this corpus has 300 documents, where we took a sample of 100 documents to realize our tests.

With the aid of the author of the thesis [24] was possible compare the degree of similarity between the summaries of our system against his system, since he provided us the corpus disasters and the summaries obtained by his system. We have to mention that the system of the thesis [24] is a system dedicated to extract specialized information of natural disasters reason why the used metrics were adapted essentially to extract certain kind of information in this type of documents, and our system is not specialized, it means is of general propose, in addition, our system is designed to always extract 5 sentences, the same that the extract given by the experts. In the corpus disasters the extracts given by the experts did not have a minimum or maximum length at the same that the extracts given by the system of the thesis [24].



GRAPHIC 1. Similarity degree between the extracts

All the previous did that the performance of similarity of our system against the extract of the experts be less. In Table 2 at the same that in the Graph 1, we can observe the degree of similarity between the summaries of both systems against the extract of the experts, observing that the degree of similarity of the extract generated by the system of the thesis [24] against the expert of the corpus disasters has better results 0.811308 against the 0.680696 of our system and the degree of similarity between the extracts generated by both systems is 0.512010, also is show the similarity degree when the sentences with anaphoras are added to the extract, in all the cases the results were less.

Table 3 shows the results of the thresholds obtained for each metric after analyzed the 100 documents that conform ours Corpus, is to say, this is the threshold that is considered important for the decision if the metric is or not used to evaluate the sentences of the document. The most used metrics that were selected by the system were: Centroid, SentencesCompare, SmallSentences, follow by TransitionPoint. The less used metrics were: ImportantWords and FisrtsLastsentences.

Table 3. Results of the obtained threshold

Metrics	General Average of the Thresholds
Centroide	0.949328
FirstLastSentences	0.249236
ImportantWords	0.090041
ProperName	0.515761
SentenceCompare	0.647503
SentencePosition	0.797081
SmallSentences	0.900797
TitleCompare	0.407710
TransitionPoint	0.837108

6 Conclusion

In this article we present a way to obtain automatic summaries by extraction in the Spanish language. The obtained results were compared against the extracts of documents made by one of the three experts and also we compare the obtained extract of the system with and without anaphoras between them.

The results obtained by the automatic summaries by extraction using the techniques that were shown in the article, have obtained favorable results, even better than the shown in the article [19]. It is important say that the results obtained for the VP by the TP [19] are good, since that algorithm obtained around 50% of the sentences of the summaries given by the experts, and in spite of the previous the new system recovered approximately 60% of the sentences, improving the effectiveness in the automatic summaries by extraction, in addition in this new article is included a study of anaphoras, with this it is tried to give a major coherence to the obtained summaries, the degree of similarity between the extract without anaphoras and the extract with anaphoras is very high, 98%, this means that after including the anaphoras did not manage to recover a major amount of important sentences.

To validate again the quality of the obtained summaries, we obtained the original CORPUS and the extracts which were used to worked in [24]. The summaries obtained by that system and ours had a level of similarity of 50%, this was because the system developed in [24] did not necessarily gives back 5 sentences, as our system did, and his metrics were adapted to keywords within the area of natural disasters, aspect that our system does not consider, so we intended to do a more general system.

The obtained results are encouraging, although the similarity levels were not so high, this does not affect the obtained results, because the way of make the summaries depends so much in the dominion of Spanish language by the person that make them.

Nowadays our group of investigation still working in the improvement of the algorithms already proven as well as new algorithms to also continue obtaining still better results in the obtaining of automatic summaries non just by extraction but by abstraction for the sake of obtaining greater precision.

References

1. Diccionario Enciclopédico de la Lengua Castellana, Ed. Codex, Buenos Aires, 1968
2. Real Academia Española. http://www.rae.es
3. Baeza-Yates R., Ribeiro-Neto B., (eds.), Modern Information Retrieval. ACM Press, New York, 1999.
4. Bueno C., Tesis de Maestría: "Métodos para la Generación de Extractos mediante el uso del Párrafo Virtual", Benemérita Universidad Autónoma de Puebla Facultad de Ciencias de la Computación, Otoño 2005, pp. 8-13.
5. Guha Sudipto, Rastogi Rajaeev, Shim Kyuseok, "ROCK, A Robust Clustering Algorithm for Categorical Attributes", Information Systems, 2000, Vol. 25, No. 5, pp. 345-366.
6. Leal L., Vilariño D., López F., Jiménez H., "The Virtual Paragraph as a Retrieval Information Technique Implanted in Mobile Agents", Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation CIMCA 2006, Sydney Australia, Nov. 2006, ISBN 0-7695-2731-0.
7. Reyes B., Moyotl E., Jiménez H., Tesis de Licenciatura "Reducción de términos índice usando el punto de transición, Facultad de Ciencias de la Computación, BUAP.
8. Urbizagástegui R., "Las posibilidades de la Ley de Zipf en la indización automática", Reporte de la Universidad de California Riverside, 1999.
9. MINITAB, http://www.minitab.com/
10. KUPIEC J., PEDERSEN J., CHEN F., A Trainable Document Summarizer, 2005.

11. Mateo P. L., González J.C., Villena J., Martínez J.L., Un sistema para resumen automático de textos en castellano., DAEDALUS, S.A., Madrid, España.
12. ROJAS F., JIMENEZ H., PINTO D., LOPEZ A., Dimensionality Reduction for Information Retrieval, 2006.
13. GUERRA A., Aprendizaje Automático: Clasificación, páginas 6-8, 2004.
14. Harabagiu S., MOLDOVAN D., CLARK Christine, BOWDEN M., HICKL A., WANG P., Employing Two Question Answering Systems in TREC-2005, 2005.
15. ZAJIC D., DORR B., SCHWARTZ R., Automatic Headline Generation for Newspaper Stories, 2002.
16. Karamuftuoglu M., Approach to Summarisation Based on Lexical Bonds, 2002.
17. TEUFEL S., MOENS M., Sentence extraction as a classification task. Workshop 'Intelligent and scalable Text summarization', 1997.
18. EDMUNSON. New Methods in Automatic Extracting. Journal of the Association for Computing Machinery, páginas 264-285, 1969.
19. MARQUEZ J. A., RENDON P. R., RODRIGUEZ R., VILARIÑO D., BELTRAN B., Comparación de dos métodos para la obtención de resúmenes automáticos, CIINDET 2007, ISBN-968-9152-00-9.
20. www.hipertexto.info/documentos/resumen.htm
21. Carcedo Elena F., "Los Géneros y su práctica", Ed. Textos UAP, 2003, pág. 57.
22. Moldovan D., Clark C., Harabagiu S. Temporal Context Representation and Reasoning, 2005.
23. Sidorov Grigori, Olivas Zazueta Omar, Resolución de anáfora pronominal para el español usando el método de conocimiento limitado, CIC, IPN, México D.F.
24. Villatoro Tello Esaú, "Generación Automática de Resúmenes de Múltiples Documentos", Tesis de Maestría del INAOE, Febrero 2007, Tonantzintla, Puebla.