Text Extraction: a Corpus-based Approach

Salazar, H.; Jiménez, H. & Pinto, D.
Facultad de Ciencias de la Computación
B. Universidad Autónoma de Puebla
C.U. 72570, Puebla, México
hilariocam@yahoo.com.mx, hjimenez@fcfm.buap.mx, dpinto@cs.buap.mx

Abstract— **In this paper we present an approach based on a corpus to automatically obtain an extract of a text. Our proposal is an inexpensive method that takes into account the semantic content of texts and uses a raw corpus as the only linguistic resource. The method is supported by the idea of sense relationships associated to a word. This method obtains the most representative sentences from a text. Such sentences were independently evaluated by four judges. The method's performance with respect to human judgment encourages the continuous development of this approach in order to carry out the text extraction task.**

Keywords: summarization, sentence extraction, corpus-based approach.

I. INTRODUCTION

The disproportion between the growth of textual information online and the tools for exploiting it entails the necessity to look for new methods that deal with the representation of and access to high volumes of text. To solve this problem several, approaches have been developed such as information extraction, which a way to provide structure to the content of a text. Summarization is perhaps a technique that may help other techniques because, in some cases, it makes the manual analysis possible. For summarization, it is therefore necessary to obtain high precision and availability in different domains. There are many methods about summarization, such as methods that use cue words [1], methods that are supported by the structure and indicators [2], and methods that are based on IR techniques [3], [4] or machine learning [5]. Such methods face a dilemma: to go through an expensive linguistic resource and thus take advantage of the semantic content of a text, or to go through some type of heuristics whose performance has been tested in certain contexts (such as Information Retrieval). In the automatic text summarization process, it is important to identify the sentences that may be used to generate the corresponding abstract, namely the extract. In the present work, we deal with the extraction process, taking the most representative sentences of a source text.

The following section presents some works related to our approach. How the method works and the performed test are explained in Section III and Section IV, respectively. At the end, we provide the conclusion of this work.

This work was partially supported by CoNaCyT grant Nr. I39186A and SEP-SESIC-DGES-VIEP grant Nr. III09G02.

II. SENSE RELATIONSHIP

In this paper we are proposing an inexpensive method that takes into account the semantic content of texts. The method is supported by the idea of sense relationships attached to a word[6]. Certainly, we use an approximation to sense relationships because we are using no more linguistics resources other than a corpus. In literature there have been works that analyze the sense relationships of a word such as the first-order terms [7] (terms that co-occur in the context of the word) and some very successful applications as the thesauri construction [8], [9], [10]. A further statement on the sense relationship appears in [11] where the nominal phrase is analyzed.

III. USING SENSES ON SENTENCE EXTRACTION

D. Marcu [12] proposed a method to compare the abstract of a paper with some sentences of the source text using a similarity function. We follow this method in part; the main differences are that Marcu's method does not use a corpus and that it starts from pairs abstract-text. We used the contexts of a word w extracted from a corpus to represent w, specifically the sentences as contexts. Our approach is simple: first, we represent each one of the sentences with the first-order terms of its components (words). Then, we find the similarity between a sentence and the whole text except by z (the complement of z). The more similarity between a sentence z and its complement, the more representative is z of the text.

In order to obtain the extract of a text automatically, the system we developed executes three steps: preprocessing, text representation and text extraction, which are described in the following subsections.

III.1. Preprocessing

The preprocessing stage is applied over both the input-text T and the corpus-used C. The goal of this stage is to identify every sentence that makes up the text and its stemmed words. The algorithm is simple and does every one of the next four steps sequentially:

1) Tokens identification.
2) Stopwords elimination.

3) Stemming.
4) Sentence segmentation.

After applying every step to the text T as well as to the corpus, the text obtained is called T_1 and C_1 respectively.

III.2. Text Representation

The vocabulary of a text is the entire body of words that appears in that text. The sense-based representation of a word w entails considering a corpus that provides the contexts of w. To each word of vocabulary V, we associate all the context of the corpus where the word occurs. Thus V may be expressed by:

$$V = \{(x, y) | x, y \text{ occurs in } S, \text{ and } S \text{ is in } C_1\}. \tag{1}$$

We conceived a sentence as a tuple, $S = (x_1, \ldots, x_k)$ (whitout repetitions), as well as the text, $T_1 = (S_1, S_2, \ldots, S_n)$. Given the text T_1 its representation is

$$T_2 = (\text{RepS}(S_1, V), \ldots, \text{RepS}(S_n, V)). \tag{2}$$

where

$$\text{RepS}(S, V) = (\text{Rep}(x_1, V), \ldots, \text{Rep}(x_k, V)). \tag{3}$$

and

$$\text{Rep}(x, V) = \{y | (x, y) \in V\}. \tag{4}$$

The algorithm that does text representation is in $O(k)$, where k is the number of words of the text to be extracted.

III.3. Text Extraction

This module uses a similarity function in order to sort every sentence S_i from T_2 in decreasing order, based on its similarity with the same text T_2. The similarity function used here is a simplification of the Jaccard similarity function:

$$sim(S_i, \bar{S}_i) = \#(S_i \cap \bar{S}_i). \tag{5}$$

where $\bar{S}_i$ is the complement of S_i in T_2 (i.e. T_2 without S_i). If T_2 is the tuple $(S_i)_i$, the ranking of sentences is performed by

$$T_3 = sort((sim(S_i, \bar{S}_i))_i, T). \tag{6}$$

Thus, T_3 is composed by the original sentences sorted by its similarity score. The complexity of this procedure is in $O(n^2)$, where n is the number of sentences represented by its sense relationships. In our test we used an arbitrary threshold of 5 to select the most representative sentences of the text.

IV. EXPERIMENTS

In our experiments, we used a corpus to determine sense relationships of terms. The corpus has defined subjects that allows us to approximately determine the sense relationships. The gathered corpus and the dataset used are described below as well as the experimental results used to check the performance of our proposal.

Subject	Num. of Docs	%
Justice	10	10.40
Culture	7	7.30
Politics	25	26.00
Society	23	24.00
Government	9	9.40
Science	11	11.50
Technology	1	1.00
Religion	2	2.10
Economy	8	8.30

TABLE I
PERCENTAGE BY SUBJECT.

Doc	Size(Kb)	Subject	Words	Sentences
1	4.4	Economy	673	25
2	4.7	Culture	614	25
3	9.3	Justice	1,379	50
4	8.4	Politics	1,223	49
5	3.8	Society	537	19
6	4.5	Society	646	25
7	4.5	Culture	644	25
8	2.5	Economy	347	16
9	5.2	Justice	731	26
10	6.1	Politics	802	21

TABLE II
DATASET SUBJECT DISTRIBUTION.

IV.1. Corpus

The gathered corpus is composed by 96 processed documents, related to the following subjects: politics, education, religion, economy, justice, culture, government, society, science and technology. The vocabulary has 22,201 stemmed terms and 4,294 sentences. Certainly, we could have taken a corpus of a determined domain; however in this experiment we wanted to know the behavior of the representation in general, as well the influence of a heterogeneous corpus in this task. The percentage of every subject as well as the number of documents by subject can be seen in Table I.

IV.2. Dataset

In order to verify the performance of the algorithm, we used ten documents with different subjects, all of which were related to the subjects used at the corpus. A better description of every document is given in Table II.

IV.3. Experimental Results

Every document of the dataset was given to four judges, so that they could dictate a judgment regarding the most representative sentences. The task of the judges was to read each

Subject	Hard Criterion	Soft Criterion
Culture	0.44	0.64
Economy	0.48	0.57
Justice	0.16	0.33
Politics	0.60	0.68
Society	0.64	0.80

TABLE III

AVERAGE EVALUATION BY SUBJECT.

document and select the five most representative sentences of each one. From the sentences selected by the judges, we considered the two top sentences as Very Representative (VR), the following two sentences as Representative (R) and the last one sentence as Sufficiently Representative (SR). As a result, we have defined three classes: VR, R and SR.

We used two criteria called "hard" and "soft" criteria, they are defined as follows:

Hard criterion: The hard criterion assigns a value of 1 if the system's and the judge's sentences fall in the same class.

Soft criterion: The hard criterion assigns a value of 1 if the system's and the judge's sentences fall in the same class. Furthermore, if the system's and the judge's judgments are on close neighbor classes (VR and R or R and SR), the value was 2/3; for far neighbor classes it was 1/3 (VR and SR).

The soft criterion was inspired by the difficulty humans have regarding the process of classifying objects into different categories.

Answers given by the algorithm were compared with the judgment of every judge using hard and soft criteria; the accuracy of these results are shown in Table III.

V. SUMMARY AND FUTURE WORK

We have presented a method to obtain the most representative sentences from a text. These sentences were independently evaluated by four judges. The results obtained were similar to what we had expected. It was possible to observe that the higher percentage of results was obtained with the subjects that also have the higher percentage in the corpus (politics and society). In spite of the small corpus used here, the results of the system, compared with the judgments, encourage the use of sense-based representation.

We are planning to compare this method with other methods such as the IR-based method, like [3]. Furthermore, it would be interesting to compare a set of indexes of a document to the words that appear in the selected sentences that provide this method. In addition, we think it is convenient to represent one more semantic unit as nominal phrases instead of only as words [11]. Finally, we would like to analyze the behavior of the method with different corpora of a single domain.

REFERENCES

[1] F. C. Johnson, C.D. Paice, W. J. Black & A. P. Neal: "The application of linguistic processing to automatic abstract generation", *SIGIR*, 1994.

[2] Gustavo Crispino, Jean-Luc Minel & Javier Couto: "Contexto: Una plataforma para la extracción de información y el resumen automático de textos", *Proceedings of the 2nd. Workshop on Spanish Processing and Language Technologies*, pp. 153-157, Universidad de Jaén, España, 2001.

[3] I. Acero, M. Alcojor, A. Díaz, J.M. Gómez & M. Maña: "Generación automática de resúmenes personalizados, *Procesamiento de Lenguaje Natural* (27), pp. 281-287, SEPLN 2001.

[4] Gerard Salton, James Allan & Amit Singhal: "Automatic text decomposition and structuring", *Information Processing and Management*, V. 32, No. 2, pp. 127-138, Elsevier, 1996.

[5] Maher Joana & Abdel majid Ben H.: "Automatic text summarization of scientific articles based on classification of extract's population, *Lecture Notes in Computer Science* 2588, A. Gelbukh (Ed.), pp. 623-634, Springer, 2003.

[6] Lyons, J.: *Semantics*, Cambridge University Press, 1977.

[7] G. Ruge: Combining corpus linguistics and human memory models for automatic term association, Text Information Retrieval, T. Strzalkowski (Ed.), Kluwer, 1999.

[8] Grefenstette, Gregory: "Automatic thesaurus generation from raw text using knowledge-poor techniques", *Xerox report*, Grenoble Lab., 1995.

[9] Grefenstette, Gregory: "Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches", *SIGLEX/ACL, Workshop on Acquisition of Lexical Knowledge from Text* Columbus, OH, 1993.

[10] Varaschin Gasperin, C. & Strube de Lima, V.L.: "Experiment on extracting semantic relations from syntactic relation", *Lecture Notes in Computer Science*, Vol. 2588, Springer, 2003.

[11] García, J.F.: "Estructura conceptual y comunicación", *Dimensión antropológica*, Año 2, Vol. 3, pp. 75-84, México, 1995.

[12] Daniel Marcu: "The Automatic construction of large-scale corpora for summarization research", *ACM-SIGIR 99*, pp. 137-144, 1999.