A New Cross-Lingua Automatic Summerization Approach Based on Textual Energy

Josue-Antonio Careaga-Moya[1], Alfonso Medina-Urrea[2], Juan-Manuel Torres-Moreno[3]

[1]Grupo de Ingeniería Lingüística, II-UNAM – jcareagam@iingen.unam.mx
[2] Grupo de Ingeniería Lingüística, II-UNAM – amedinau@iingen.unam.mx
[3]Laboratoire Informatique D'Avignon, LIA – juan-manuel.torres@univ-avignon.fr

Abstract

The main target of cross-language summarization is to generate a summary in a different language from the language of the source document or documents. In this paper, it is proposed a Textual Energy approach to mono-document summarization plus the use of a Machine Translation online system to translate the source file from English into Spanish. Proficiency of the system was measured with FRESA[11] framework and compared with baseline summaries generated at different percentages.

Keywords: cross-language summarization, textual energy, machine translation, natural language processing, information extraction.

1. Introduction

Searching for information available on the web requires queries to be specific in order to obtain results which match user interests. Some of the factors that influence this need are: the diversity of information sources, the various languages in which such information is available, and the length of documents that contain it. For these reasons, search engines able to predict and make easier to users the search task, as well as dictionaries, automatic translation tools, summarizers, among others, have become high demanded applications.

This work is focused on summarizing news translated from English to Spanish. The aim is to make information more accessible and thus to increase the information (news) perspective of Spanish readers by simplifying the task of reading, translating and then summarizing texts.

The automatic summarization task can be accomplished by two types of methods: statistical and linguistic techniques (da Cunha *et al.*, 2009). Thus, for each type of technique there are methods like clustering, graph implementation, Bayesian models, as well as those based on word/

1 FRESA-FRamework for Evaluating Summaries Automatically http://daniel.iut.univ-metz.fr/~LIA_TALNE/FRESA/

sentence location, lexical chains or discourse structure, among others. Another classification refers to how the summary is generated, whether extractive or generative.

Extractive and generative methods for automatic summarization have received much attention. Nevertheless it is known that natural language generation is not an easy task, because generation of text that makes sense is such a hard work. Statistical techniques are involved regardless of the chosen method.

For instance, extractive techniques based on word counting, location and cue phrases are still considered basic techniques for summary generation. This work is related to an extractive technique as well, considering the top of the ranked sentences by Textual Energy scores from the input document.

Also, cross-language summarization can be approached in different ways, such as working with mono-document, multi-document, bilingual, and multi-lingual, among others. Regarding these, there are variations that consider source documents written in more than one language, different documents written in different languages, or the possibility to generate the resulting summary in more than one language.

With respect to Machine Translation, there is a general classification of this task comprising: the linguistic method, the statistical one and the hybrid one, which combines both. The main problem for MT is to keep the input text as it is once the translation has been done.

The main idea on this study is to measure the time that the translation process of each utterance or expression takes and multiply these measurements by the textual energy scores matrix, which contains textual energy score of each utterance. Needless to say, all these values are stored in two matrices: one for the time measurements and one for the textual energy scores. The result of multiplying these matrices is reflected into the results of the system itself.

2. Related Work

2.1 Automatic Summarization

There are several research projects that involve automatic summarization, based on the existing methods for generating automatic summaries.

Automatic summarization by clustering and extracting information has been one of the most implemented techniques since it brings tools that help users to extract and measure which information is actually relevant by sentence ranking through word frequency in the text. An example of this is the work made about mouse gene information (Yang, 2007) which uses information extraction and clustering for summary generation.

CORTEX (Torres-Moreno, 2001) uses frequency of word appearance by combining several metrics obtained from statistical algorithms and information from a vectorial representation of the document. It considers the presence-absence of phrases or sentences and a vocabulary of terms. Some of the metrics this system uses are Hamming matrix, entropy, angle between title and phrases, among others. These metrics are scored to generate the wanted summary.

2.2 Cross-language summarization

As it was mentioned, depending on how the source document is written, there are different methods to work with cross-language summarization. In the multi-document and multi-language summarization task, the Columbia Newsblaster system (Evans, 2004) shows a robust system, which crawls, translates, clusters, summarizes and classifies documents (news) written in different languages from the web.

The Columbia Newsblaster system generates a web page, which concentrates all the information that was translated by the Systran system and summarized by the Columbia Summarizer. The latter is based on the similarity of the documents in the cluster.

Another approach for cross-language summarization is the Graph-based approach (Boudin, 2011) that considers translation quality of sentences during the sentence selection process by a supervised learning approach. The summarization process is done in two steps: scoring of each sentence of the document and then, selecting the top ranked sentences to be included in the summary. The translation process is added as a previous step and is done using the Google translator system.

3. Method

Our study proposes a cross-language summarization system that uses textual energy and time measurement of translation. This makes final summaries of news more reliable. The automatic summarization method using textual energy implies takes a statistical physics inspired method and combines it with a Vector Space Model VSM and neural networks.

The ENERTEX method (Fernández, 2007) considers the words in a text as sets of units, which interact among each other and are affected by the field each one of them generates. Thus, each word gets a score according to its textual energy.

This approach considers the time of translation of each utterance. It then generates a textual energy matrix which will contribute to the generation of the summary. The evaluation of the system was conducted with the FRESA framework, considering baseline summaries created automatically for different percentages of the original texts.

Our method's system architecture is shown in Figure 1.

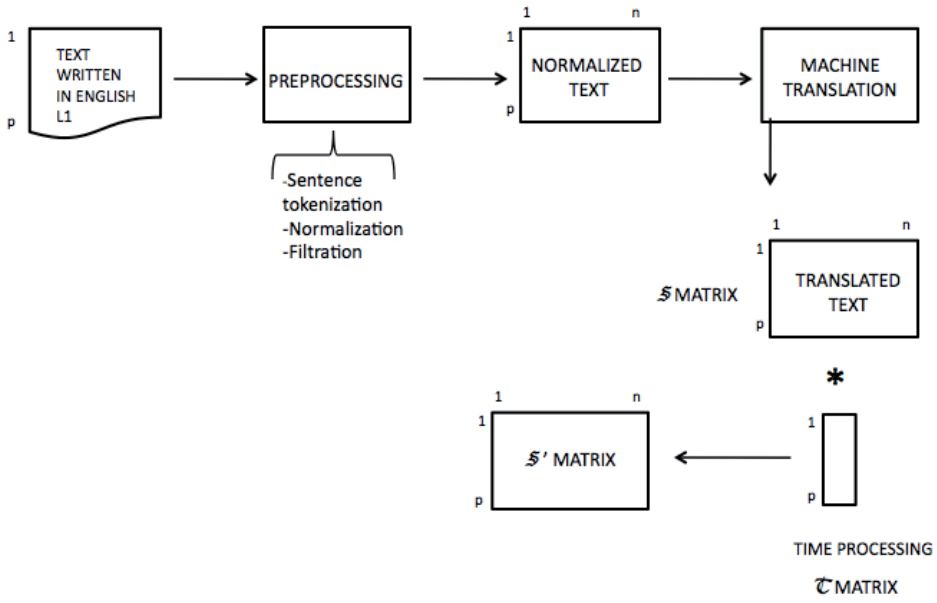

Figure 1. System Architecture

The input to our system is a news document in English. It is segmented into utterances (expressions between periods) using the NLTK Chunker for Python, so that they can be represented by vectors using the VSM. Then, these are normalized and filtered so that they can be fed to a machine translation system.[2]

The translation process generates a matrix S of utterances (rows) and the words of those utterances (columns). Also, we get a column matrix T containing the processing time for each utterance (time measurement of its translation process). To associate utterance processing time to the utterance itself, both these matrixes (S and T) are multiplied.

$$S' = S * T^{3}$$

Once matrix S' is obtained, it permits us to calculate an E matrix portraying the textual energy of the text. This E matrix is obtained by squaring the product of S' times its transposed matrix.

$$E = (S' \times S'^{T})^{2}$$

When matrix E is already formed, we sum up the rows and use these values to obtain a new column matrix of relevance measurements to rank utterances by importance.

$$| \Sigma E |$$

2 http://reverso.net

3 The operation can be also expressed as S'=S* diag(T) in a matlab syntax.

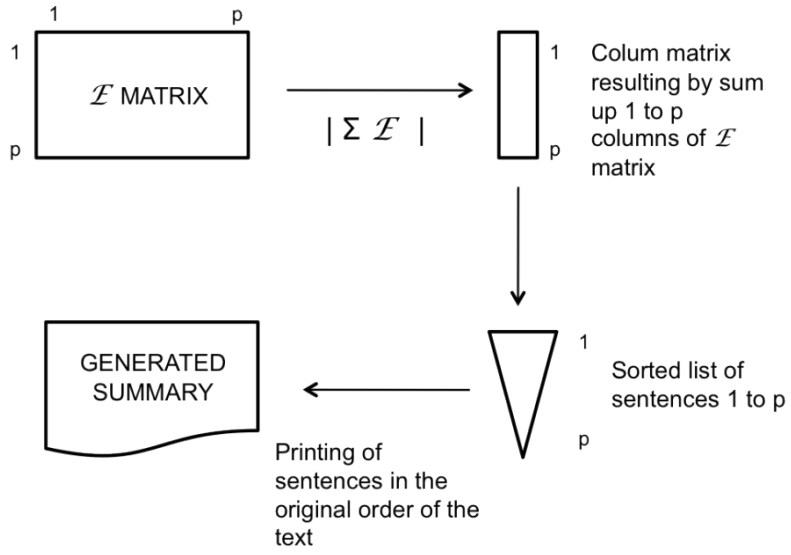

Figure 2. Summary generation from E matrix.

The sorted list obtained (column matrix E) is the one that is used to take the number of utterances that the user wants the summary to be made of, If the user requires an extract of 10% of the input document, the system will take those utterances that conform 10% of the original text and they will be inserted into the output file sorted respect to original appearance in the source document.

Finally, to evaluate the proficiency of our system we took each of the generated summaries and introduced them into the FRESA system. And then we compare the result of the divergences obtained with the baseline summary that the system generates itself. On next the section, the FRESA evaluation functionality is described and the results of our system are shown.

4. Results

From the twenty-five news that were recompiled from the CNN website, two of them were too short and stuck the system in a loop error. Therefore, those documents were not taken into account to be processed.

We used the remaining set of news, which was processed by our system. Since we did not have baseline summaries provided by the news authors, we generated automatically baseline summaries by gathering the first sentences of the documents at different percentages: 10%, 15% and 20%. Each of these baselines was evaluated with the same FRESA framework as well. This way, we had six kinds of generated summaries; three of them were the baseline ones (at 10%, 15% and 20%) and the other three came from the ENERTEX system (at the same percentages).

The FRESA framework takes the input and generates a summary of it. This summary is compared to the summary submitted (made by our system) and then, statistical divergences are

calculated between them. Those divergences are: Kullback-Leibler (KL)[4] and Jensen-Shanon (JS)[5]. The smaller these divergences, the better quality of summary is obtained.

As a particular analysis of the proficiency of the proposed system, we present two tables with divergences calculated for one particular document. Later, we present graphics containing the means of both divergences among the six kinds of generated summaries.

Table 1 shows the results of JS divergences obtained by FRESA compared with baseline summaries of the same text generated at 10%, 15% and 20% of the original content of the document.

Summary Type	10%	15%	20%
Baseline first sentences	2.46236	2.46236	2.46236
ENERTEX System	2.38968	2.57715	2.57715

Table 1. Jensen-Shanon divergence results evaluated by FRESA

Looking at this table we realize that the ENERTEX summarization proposal is better than the baseline generated summary of the first sentences comprising 10% of the text. Regarding the other summary percentages, we can see that ENERTEX approach is over the baseline by 0.11, so that ENERTEX was not good enough. These results are from one of the shortest documents.

Table 2 shows the KL divergences calculated by FRESA compared with baselines summaries at 10, 15 and 20%. Here we can see that values of KL divergence are lower than the baseline summary obtained for 10% of the original text. These values are from the same file at the above table.

Summary Type	10%	15%	20%
Baseline first sentences	11.59062	11.59062	11.59062
ENERTEX System	11.48663	12.03997	12.03997

Table 2. Kullback-Leibler divergence results evaluated by FRESA

By checking KL divergences we can see again that, in the other summaries (those conformed by 15% and 20% of the original text), ENERTEX did not performed as well, showing a difference of 0.44 against the baseline summary generated.

In a more general analysis of our system proficiency (ENERTEX), we have the following graphics to show the behavior of the system among all the documents by calculating and representing the mean of each divergence on different graphs.

4 Kullback-Liebler Divergence recall: $D_{KL}(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$

5 Jensen-Shanon Divergence is recall: $JSD(P\|Q) = \frac{1}{2} D(P\|M) + \frac{1}{2} D(Q\|M)$

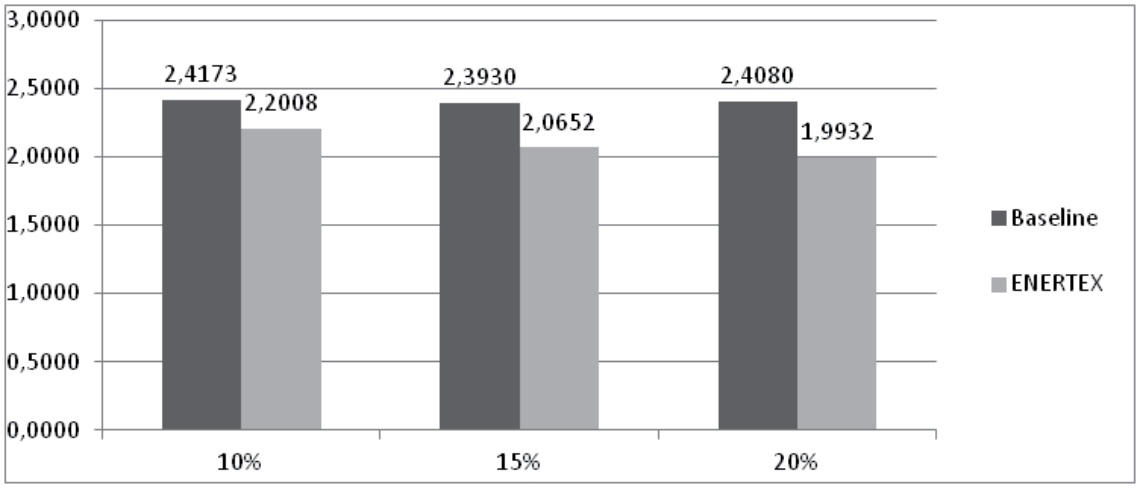

Figure 3. Jensen – Shanon divergence mean analysis with ENERTEX system.

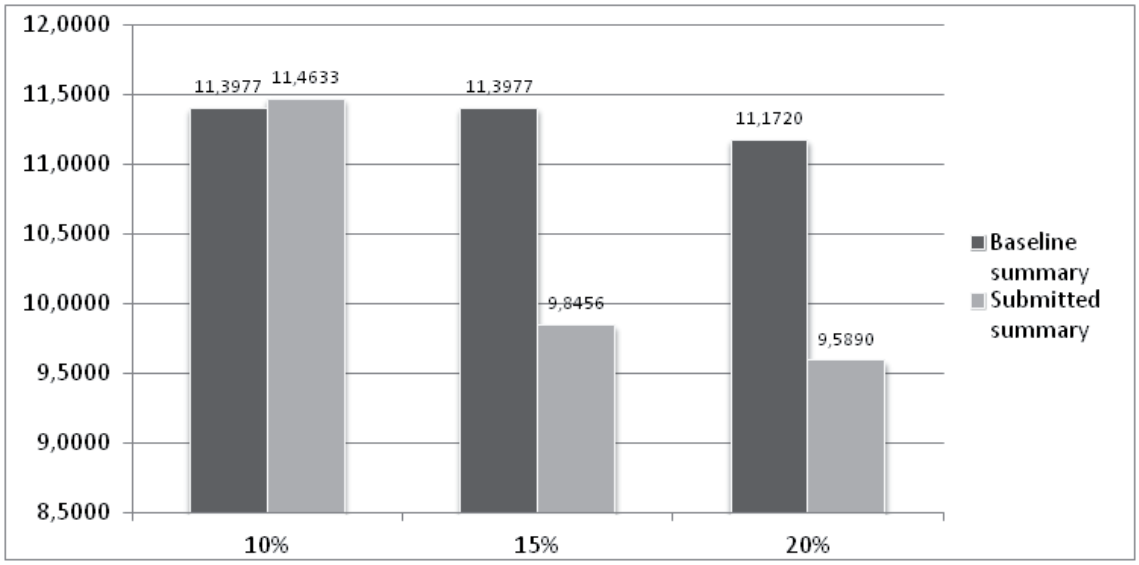

Figure 4. Kullback – Leibler divergence mean analysis with ENERTEX system.

From both previous graphics, we can see that with short summaries, for instance at 10% of the original text, the ENERTEX approach is not under the baseline divergence, which implies that it is not the best summarization approach for those kinds of summaries. But the result is still coherent and relevant for 15% and 20% summaries. Actually, the difference is much higher than 0.07, which is the difference on the 10% case; difference for 15% is 1.54.

In the case of the baseline summaries (generated from the first sentences of the documents) we can see on the graphs below (figures 5 and 6) that this method performed well, because in every percentage considered (10%, 15% and 20%) the values of JS and KL divergences are under the baselines generated by the FRESA system values and these are, as it was mentioned, the expected results for a good summarizer system.

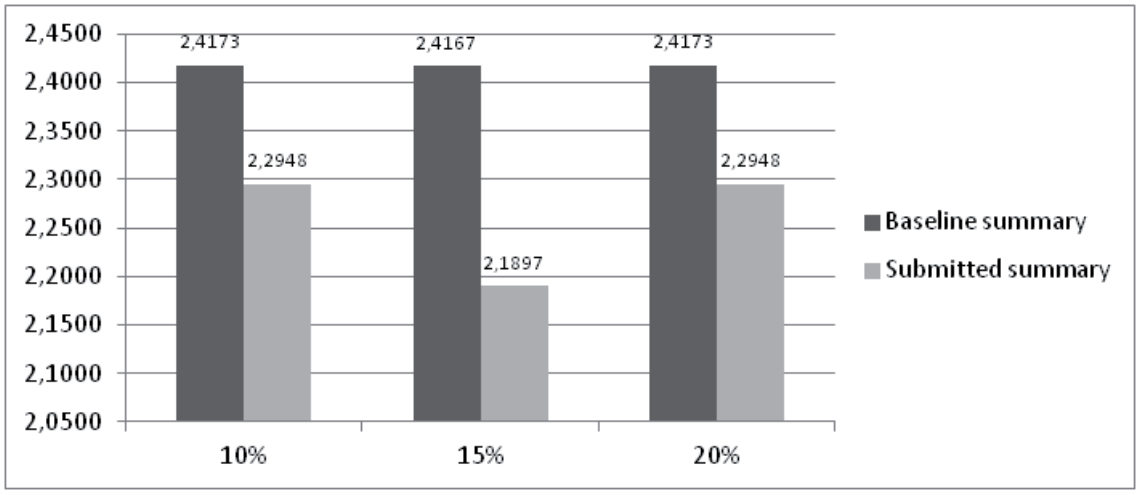

Figure 5. Jensen – Shanon divergence mean analysis of baseline summaries

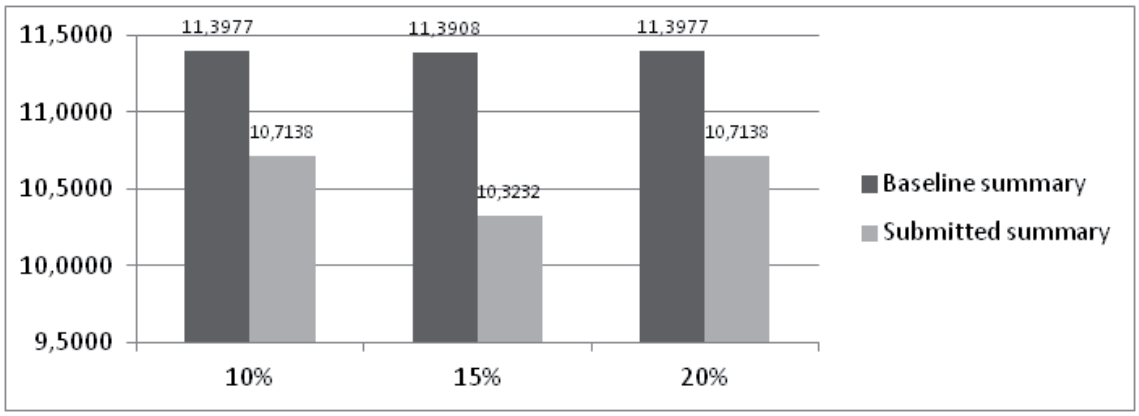

Figure 6. Kullback - Leibler divergence mean analysis of baseline summaries.

5. Conclusions

As results show, the proposed system is capable to generate better summaries when: (1) the source documents are not too short and (2) when the desired summary is asked to be formed by 15% or 20% of the original data. So, these reasons are correlated because of the statistic technique that textual energy involves.

More specifically, when means of the divergence were calculated we noted the above observations when comparing summaries of 10% and 15% and 20% of the original text. When comparing to baseline summary of FRESA, both divergences (JS and KL), the values were lower for at least 1.59 units on the means.

It can be concluded that our cross-language summarization approach, which is based on textual energy, is a good technique according to the results of the generated summaries. These results were good enough based on the comparison made with the baselines summaries, with the exception of those which were summaries from short documents.

In general, cross-language summarization will deal with machine translation (MT) difficulties. Inspite of considerable obstacles, the current work and research on the subject is quite extensive and complex.

However, more techniques can be combined in order to improve MT and, consequently, to obtain better results on the cross-language summarization task. Furthermore, the chance remains to innovate by creating different combinations of existing approaches.

For further work there is another aim to analyze how English speakers cultural and geographical situation modifies the way they write news and the way they actually perceive them.

References

Boudin, F., Huet, S. and Torres-Moreno, J. 2011. A Graph-based Approach to Cross-language Multi-document Summarization. Polibitis 43:113-118

da Cunha, I., Torres-Moreno, J., Velázquez-Morales, P. and Vivaldi, J. 2009. Un algoritmo lingüístico-estadístico para resumen automático de textos especializados. LinguaMÁTICA 2:67-80.

Evans, D., Klavans, J. and Mckeown, K. 2004. Columbia Newsblaster: Multilingual News Summarization on the Web. Association for Computational Linguistics. 1-4.

Fernandez, S., SanJuan, E., and Juan-Manuel, T. 2007. Textual Energy of Associative Memories: performants applications of Enertex algorithm in text summarization and topic segmentation. MICAI 2007

Luhn, H.P. 1959. The automatic creation of Literature abstracts. IBM Journal of research and development, 2(2).

Mani, I. and M.T. Maybury. 1999. Advances in automatic text summarization. Cambridge, MA: MIT Press.

Ogden, W., Cowie, J., Davis, M., Ludovik, E., Molina-Salgado, H. and Shin, H. Getting Information from Documents You Cannot Read: An Interactive Cross-Language Text Retrieval and Summarization System.

Torres-Moreno, J.M., Velázquez-Morales, P. and Meunier. J.G. 2001. Cortex : un algorithme pour la condensation automatique des textes. Proc. of ARCo 2001, pp. 65–75.

Yang, J., Cohen, A. and Hersh, W. 2007. Automatic Summarization of Mouse Gene Information by Clustering and Sentence Extraction from MEDLINE Abstracts. AMIA Annu Symp Proc. 2007; 2007: 831–835.