Extractive Summarization Based on Word Information and Sentence Position

Carlos Méndez Cruz and Alfonso Medina Urrea

GIL IINGEN UNAM, Apartado Postal 70-472,
04510 Coyoacán, DF, Mexico
{cmendezc, amedinau}@iingen.unam.mx

Abstract. This paper describes an unsupervised experiment of automatic summarization. The idea is to rate each sentence of a document according to the information content of its graphical words. Also, as a minimal measure of document structure, we added a sentence position coefficient.

1 Introduction

Plenty of research has been conducted in the field of automatic summarization. The need for such methods has motivated the exploration of many approaches which reflect the field's complexity. Many aspects of texts must be considered, such as word frequency, document, paragraph and sentence structure, topic and focus structure, information content, etc. Many of these approaches are based on the idea that the greater number of times a linguistic structure or part of it occurs (a word, phrase, sentence, etc.), the more attention the reader will pay to it except when it is a function or grammatical word. That is, a document's sentences receive different levels of attention by human readers. Current interests among researchers are multi-document summarization [1, 2], the application of artificial intelligence methods such as genetic algorithms [2], the use of lexical chains and web resources such as WordNet [3].

Since we are currently developing an open, Spanish language corpus on engineering (CLI) to be available on the Internet [4], we are exploring some summarization techniques to apply to it. The main criteria for this very first experiment was to avoid the heavy techniques that have been and can be developed if one takes into account the complexity of document, paragraph, sentence, and word structure. Thus, we opted for an unsupervised approach based on simple information content measurements that could conceivably be applied to other languages. Actually, information content estimates are typically used for a wide variety of unsupervised tasks. And in fact, some experiments have explored the notions of information content and entropy models for some aspect or another of automatic summarization — for instance, summary evaluation or reductive transformation [5, 6, 7]. In this paper, we will first define some basic concepts. Then, we will briefly describe our application and lastly, we will present results and evaluation strategy.

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 653–656, 2005.

2 Basic Working Concepts

A summary is a reductive transformation of a source text through content reduction by selection of what is important in that source [8]. It is well known that the main problem is to capture the important content of the source text. In general, two sorts of summaries can be produced: extracts and abstracts. The first ones are made by transferring part of the source text to what constitutes the summary. In the latter ones, it is necessary to modify the output to build a clear summary. Even though it is well known that extracts deliver a lower-quality output, we constrained this experiment to generating extracts.

It is also necessary to define what we mean by sentence, since we are dealing with an unsupervised method and thus, clause or sentence structure is really not considered as such. In short, we will here call *sentence* whatever occurs between two periods,[1] that is, the set of phrases surrounded by periods.

3 Method

A Python-NLTK program was developed, which ranks each document sentence according to an index estimated from the information content of its graphical words: $\log_2(p_i)$, where i refers to the graphical word, and p_i to its relative frequency in the document at hand (TF). Thus, in order to obtain a ranking index for a sentence of n words, we can simply average word information: $\frac{1}{n}\sum_{i=1}^{n}\log_2(p_i)$.

Also, we introduced a sentence position coefficient which modifies the index to give prominency to sentences occurring towards the end of the document. The idea is that the latter part of the text is more likely to present informative sentences. Thus, if we take o to be the offset or position of a sentence in a document and s to be the number of sentences in that document, then $\sqrt[25]{\frac{o}{s}}$ grows rapidly for sentences occurring at the beginning of the document and is greater for sentences occurring towards the end.

Hence, the index we used to rank each sentence combines both, word information and sentence position:

$$\frac{\sum_{i=1}^{n}\log_2(p_i)}{n} * \sqrt[25]{o/s} \tag{1}$$

Instead of using a stop list to screen out grammatical words, we used a filter which permitted us to screen function words and infrequent ones. Since function words typically contain the least information, we opted to screen out those word types with less than half of the overall information average in the document:

$$\frac{\sum_{i=1}^{t}\log_2(p_i)}{2t} \tag{2}$$

[1] We are well aware of the use of periods to signal abbreviations (in Spanish and many other languages). However, we have opted not to deal with this mainly because we are seeking unsupervisedness at this point. For our purposes, abbreviations simply cut sentences into smaller units to be ranked as eligible summary candidates.

where t is the number of word types in the document at hand. Lastly, we also filtered out words with frequency in that document of less than 3.

Based on all of this, the program selects for each targeted document the ten sentences with the highest index values in order to produce the final summary. Then, the sentences are listed according to their position in the source text and presented for evaluation.

To test this method, we selected seven documents from the CLI. Five of them were long, technical reports and two were short articles. All of them belong to different thematic areas of engineering: mechanical, electric and electronic engineering. Also, for the sake of comparison, we included two humanities texts (science-fiction and literary criticism).

4 Results and Evaluation

Summarization requires rigorous evaluation. However, the criteria for accomplishing this are so elusive, that it ends up being inevitably subjective. Our simple approach is not likely to do better than heavier approaches, but we devised a simple evaluation scheme to judge results against maximum possible scores restricted to the documents mentioned above.

In essence, we requested eight subjects to read each of the generated summaries and to write a brief text recreating the source text. Thus, they wrote a description of what they thought the source document was about. If the subject guessed the main idea of the source text, a score of 1 was registered, otherwise 0.

It is interesting to note that texts 5, 6 and 7 — which obtained the lowest scores — were the very long, technical reports with many scientific notational idiosyncracies, as well as figures and tables, whose traces appeared as part of the summaries. This made it difficult for the subjects to even read the extracts.

From these scores we can estimate the relative number of positive scores (subjects guessed the main idea of a document a total of 50 times) with respect to the possible number of positive scores (eight readers and nine documents means 72 possible positive scores): $50/72 = 0.69444$. This is a sort of precision measure, which deals with whether or not the subjects' guesses were right. However, it is also important to look at how complete the guesses were. For this, we assigned a score from 0 to 2 to each of the subjects' texts; where 0 meant much of the relevant information was missed, 1 meant some important information was omitted, and 2 no important information was missing.

The long, technical reports — texts 5, 6 and 7 — received again the lowest scores. The much shorter, technical articles and the humanities papers obtained the best scores. This second set of scores can be better appreciated if we consider the relative value of total scores (an accumulated score of 66) with respect to the maximum possible total sum of scores (twice 72). That is, $66/144 = 0.45833$. This value would be a kind of recall measure.

Although these precision and recall measures look encouraging, they constitute no appropriate criteria for comparison to other experiments. Such an

important evaluation remains to be done and will certainly require much more attention than what can be paid in this reduced space.

5 Conclusions

We have presented the results of a very basic and constrained experiment of unsupervised automatic summarization. From the evidence presented, we can conclude that information content should be further explored as a method for reductive transformation (not only summary evaluation).

This experiment can be taken further by varying the number of sentences to be included in the summary, the information content threshold for considering graphical words and sentence position coefficient. Also, we expect that the results can be much improved by including a lemmatization stage — particularly important for an inflectional language like Spanish[2] — and advancing to supervised methods which consider document, paragraph, sentence, phrase and morphological word structure.

Acknowledgments

The work reported on this paper has been supported by DGAPA PAPITT's Project IX402204.

References

1. Saggion, H., Gaizauskas, R.: Multi-document summarization by cluster/profile relevance and redundancy removal. In: DUC. (2004)
2. Jaoua, M., Ben Hamadou, A.: Automatic text summarization of scientific articles based on classification of extract's population. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing Proceedings. (2003)
3. Song, Y.I., Han, K.S., Rim, H.C.: A Term Weighting Method based on Lexical Chain for Automatic Summarization. In: Lecture Note in Computer Science. Springer (2004) 636–639
4. Medina, A., Sierra, G., Garduño, G., Méndez, C., Saldaña, R.: "CLI: An Open Linguistic Corpus for Engineering". In: Iberamia. (2004) 203–205
5. Ravindra, G., Balakrishnan, N., Ramakrishnan, K.R.: "Multi-Document Automatic Text Summarization Using Entropy Estimates". In: SOFSEM. (2004)
6. Ro, Park.H., S., Han.Y., H, Kim.T.: "Heuristic algorithms for automatic summarization of Korean texts". In: Online Proceedings ICCS/JCSS99. (1999)
7. Hovy, E.: "Text Summarization". In: The Oxford Handbook of Computational Linguistics. Oxford UP (2003) 583–598
8. Spark Jones, K.: Automatic summarizing: factors and directions. In: Advances in automatic summarization. MIT (1999) 1–15
9. Gelbukh, A., Sidorov, G.: "Morphological Analysis of Inflective Languages through Generation". Procesamiento de Lenguaje Natural (2002) 105–112

[2] There already exist suitable tools for lemmatization of Spanish words [9].