Experiments on Applying a Text Summarization System for Question Answering

Pedro Paulo Balage Filho, Vinícius Rodrigues de Uzêda,
Thiago Alexandre Salgueiro Pardo, and Maria das Graças Volpe Nunes

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo
CP 668, 13.560-970 São Carlos-SP, Brasil
http://www.nilc.icmc.usp.br
{pedrobalage,vruzeda}@gmail.com, {taspardo,gracan}@icmc.usp.br

Abstract. In this paper, we present and analyze the results of the application of a text summarization system – GistSumm – to the task of monolingual question answering at CLEF 2006 for Portuguese texts. We hypothesized that topic-oriented summarization techniques could be able to produce more accurate answers. However, our results showed that there is a big gap to be overcome for summarization to be directly applied to question-answering tasks.

Keywords: Question answering, text summarization.

1 Introduction

We present and analyze in this paper the results of the application of a summarization system to the task of monolingual Question Answering (QA) at CLEF 2006 for Portuguese texts. We aimed at assessing the performance of the system in answering questions using its topic-oriented summarization method: each question was considered the topic around which the summary should be built, hopefully containing the appropriate answer. The idea is that summarization techniques may be able to produce more accurate answers, since the irrelevant information in texts is ignored.

The system we used is GistSumm [3], a summarizer with very high precision in identifying the main idea of texts, as indicated by its participation in DUC (Document Understanding Conference) 2003 evaluation.

Two runs were submitted to CLEF. For one run, we submitted the summarization system results without any post-processing step. For the other one, we applied a simple filter that we developed for finding in the produced summary a more factual answer. The performance of both methods at CLEF was very poor, indicating that simple summarization techniques alone are not enough for question answering tasks.

The summarization system we used and the filter we developed are briefly described in the next section. Our results at CLEF are reported in Section 3.

2 The System

GistSumm is an automatic summarizer based on an extractive method, called gist-based method. In order to make GistSumm to work, the following premises must hold: (a) every

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 372 – 376, 2007.

text is built around a main idea, namely, its gist; (b) it is possible to identify in a text just one sentence that best expresses its main idea, namely, the gist sentence. Based on them, the following hypotheses underlie GistSumm methodology: (I) through simple statistics, the gist sentence or an approximation of it can be determined; (II) by means of the gist sentence, it is possible to build coherent extracts conveying the gist sentence itself and extra sentences from the source text that complement it.

The original GistSumm comprises three main processes: text segmentation, sentence ranking, and extract production. Text segmentation identifies the sentences in the text. Sentence ranking is based on the keywords method [2]: it scores each sentence of the source text by summing up the frequency of its words in the whole text. For producing summaries (i.e., generic summaries), the gist sentence is chosen as the one with the highest score. Extract production focuses on selecting other sentences from the source text to be included in the extract, based on: (a) gist correlation and (b) relevance to the overall content of the source text. A new version of GistSumm [1] is able to treat and summarize structured texts, e.g., scientific texts, producing an overall extract composed of smaller extracts of each text section. A characteristic of both systems is the possibility to produce summaries based on a specific topic suggested by the user. For this, the gist sentence is chosen as the one with the highest score in relation to the specified topic, with this correlation being measured by the cosine measure [4].

For participating at CLEF, we submitted two runs. For the first run, we simply returned, for each question, the highest scored gist sentence found in the topic-oriented summarization mode, with the question being the topic. In this manner we were searching for possible answers on the text that had better correlation with the question. For running CLEF experiments, the gist sentence was chosen independently for each section of each text in CLEF database. The best gist sentence was then selected to be the answer.

For the other run, we developed a filter for finding, for each question, a more restricted answer inside sentences previously indicated by GistSumm. These sentences were empirically set to be the 6 best scored ones in the topic-oriented summarization mode for the complete data collection, i.e., the 6 sentences in the whole text collection with the best correlations to the question. We forced the system to select sentences from different texts.

For each question, the filter works as follows:

- initially, it annotates the 6 selected sentences using a POS tagger;
- then, it tries to determine the type of the question posted by CLEF by analyzing its first words: if the question starts with “who”, then the filter knows that a person must be found; if it starts with “where”, then a place must be found; if it starts with “when”, then a time period must be found; if it starts with “how many”, then a quantity must be found; for any other case, the filter aborts its operation and returns as answer the same answer that would be given in the first run, i.e., the highest scored gist sentence;
- if the question type could be determined, then the filter performs a pattern matching process: if the question requires a date as answer, the filter will look for text spans in the 6 sentences that conforms, for example, to the pattern “month/day/year”; if a person or a place is required, than the filter will search for proper nouns (indicated by the POS tagger); if a quantity is required, the filter will search for expressions formed by numbers followed by nouns;

- if it was possible to find at least one answer in the last step, the first answer found is returned; otherwise, the answer is set to NIL, which ideally would indicate that the text collection do not contain the answer to the question.

The obtained results for both methods are reported in the next section.

3 Results and Discussion

We run our experiments on the Portuguese data collection for the following reasons: Portuguese is our native language and, thus, this enables us to better judge the results; Portuguese is one of the languages supported by GistSumm. CLEF Portuguese data collection contains news texts from Brazilian newspaper *Folha de São Paulo* and Portuguese newspaper *Público*, from years 1994 and 1995.

The QA track organizers made available 200 questions, which included the so called "factoid", "definition", and "list" questions. As defined by CLEF, factoid questions are fact-based questions, asking, for instance, for the name of a person or a location; definition questions are questions like "What/Who is X?"; list questions are questions that require a list of items as answers, for example, "Which European cities have hosted the Olympic Games?". There were different amounts of each question type in QA track: 139 factoid questions, 47 definition questions, and 12 list questions.

The main evaluation measure used by CLEF is accuracy. For this measure, human judges had to tell, for each question, if the answer was right, wrong, unsupported (i.e., the answer contains a correct information but the provided text do not support it, or the text do not originate from the data collection), inexact (the answer contains a correct information and the provided text support it, but the answer is incomplete/truncated or is longer than the minimum amount of information required), or unassessed (for the case that no judgment was provided for the answer). Some variations of the accuracy measure are the Confidence Weighted Score (CWS), the Mean Reciprocal Rank Score (MRRS) and the K1 measure. For the interested reader, we suggest to refer to CLEF evaluation guidelines for these measures definitions. Table 1 and 2 show the results for the two submitted runs.

One can see that the results for both runs are very poor. The results were slightly better for the second run, given the use of the filter, even though the 3 right answers for the non-list questions originated from NIL answers. The performance of our system can be better visualized in the example (from CLEF data) in Figure 1. In this example, the system returned the 6 best correlated sentences with the question. Although the cosine measure is efficient when we desire to identify sentences of a topic to compose a summary, we can

Table 1. Results for the first run (without filtering the results)

Accuracy			CWS	MRRS	K1
Questions Judgments	Factoid and definition	List	0	0	-0.6445
Right	0	0			
Wrong	179	9			
Unsupported	7	0			
Inexact	2	3			
Unassessed	0	0			

Table 2. Results for the second run (filtering the results)

Accuracy			CWS	MRRS	K1
Questions Judgments	Factoid and definition	List	0.0002	0.0160	-0.5134
Right	3	0			
Wrong	179	10			
Unsupported	4	0			
Inexact	2	2			
Unassessed	0	0			

observe that it does not show to be good for question answering. In this example, we can also see that the filter was not applied because the question identifier ("What is the name") wasn't in its rules.

Analyzing the performance of the system for the entire set of questions and answers submitted, we could identify some important points to be investigated in the future:

- the sentence level analysis usually carried out by summarization systems are not sufficiently fine-grained to appropriately answer questions;
- although the cosine measure may be good for building topic-oriented summaries, it is not good for searching for answers, since it tends to select shorter sentences that have more common words with the question, ignoring longer sentences that may probably contain the answer;
- although the answer could not be evidenced by the best scored sentence, an analysis of the other sentences showed that the answer could be found among the 100 better scored sentences, in general;
- the filter we developed was too naive in trying to classify the questions in only 4 types ("who", "where", "when" and "how many"); many more variations were used in CLEF.

After CLEF experiments, we believe that simple summarization techniques are not enough for question answering tasks, even if these systems are good in what they do. We believe that much more work in this direction is needed.

Question: *Como se chama a primeira mulher a escalar o Evereste sem máscara de oxigênio?* (in English, What is the name of the first woman to climb Everest without oxigen mask?)

First selected answers and their cosine measure (in relation to the question):

Cosine	Answer
0.500000	*Ou: Uma mulher como eu* (Or: A woman like me)
0.500000	*Como se chama?* (What is her name?)
0.500000	*Como se chama?* (What is her name?)
0.472456	*Em Maio deste ano, Hargreaves, de 33 anos, tinha se tornado a primeira mulher a escalar o Evereste, sozinha e sem a ajuda de oxigênio.* (In May of this year, Hargreaves, 33 years, became the first woman to climb Everest, alone and without the oxygen aid.)

Selected answer for 1st run: *Ou: Uma mulher como eu* (Or: A woman like me)
Selected answer for 2nd run: *Ou: Uma mulher como eu* (Or: A woman like me)

Fig. 1. Example of an answer submitted to CLEF

Acknowledgments

The authors are grateful to CAPES, CNPq and FAPESP for supporting this work.

References

[1] Balage Filho, P.P., Uzêda, V.R., Pardo, T.A.S., Nunes, M.G.V.: Estrutura Textual e Multiplicidade de Tópicos na Sumarização Automática: o Caso do Sistema GistSumm. Technical Report 283. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (2006)
[2] Luhn, H.: The automatic creation of literature abstracts. IBM Journal of Research and Development 2, 159–165 (1958)
[3] Pardo, T.A.S., Rino, L.H.M., Nunes, M.G.V.: GistSumm: A Summarization Tool Based on a New Extractive Method. In: Mamede, N.J., Baptista, J., Trancoso, I., Nunes, M.G.V. (eds.) PROPOR 2003. LNCS, vol. 2721, pp. 210–218. Springer, Heidelberg (2003)
[4] Salton, G.: Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, Reading (1989)