Using a Text Summarization System for Monolingual Question Answering

Pedro Paulo Balage Filho
Vinícius Rodrigues de Uzêda
Thiago Alexandre Salgueiro Pardo
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. Using the system ability to produce topic-oriented summaries, we aimed at assessing its accuracy in finding answers to the posted questions, which were used as the topics for producing the corresponding summaries. The obtained results were very poor, indicating that simple summarization techniques alone are not enough for question answering.

Categories and Subject Descriptors

H.3 [INFORMATION STORAGE AND RETRIEVAL]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; I.2 [ARTIFICIAL INTELLIGENCE]: I.2.7 Natural Language Processing

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, text summarization

1. Introduction

We present 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 system we used is GistSumm (Pardo et al., 2003), a simple 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 for evaluation. 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 trying to find in the produced summary a more factual answer. With this, we wanted to have shorter answers containing, for instance, only the name of a person or a specific date, instead of whole sentences. 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 a simple extractive method, called gist-based method. For GistSumm to work, the following premises must hold: (a) every 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 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.

GistSumm comprises three main processes: text segmentation, sentence ranking, and extract production. Text segmentation is carried out by an automatic sentence splitter called SENTER (Pardo, 2006). Sentence ranking is based on the keywords method (Luhn, 1958): it scores each sentence of the source text by summing up the frequency of its words in the whole text. For producing non-topic-oriented summaries (i.e., generic summaries), the gist sentence is chosen as the one with the highest score; for producing topic-oriented summaries, the gist sentence is the sentence with the best correlation to a specified topic, with this correlation being measured by the traditional cosine measure (Salton, 1989). Extract production focuses on selecting other sentences from the source text to include in the extract, based on: (a) gist correlation and (b) relevance to the overall content of the source text. Criterion (a) is fulfilled by simply verifying co-occurring words in the candidate sentences and the gist sentence, ensuring lexical cohesion. Criterion (b) is fulfilled by sentences whose score is above a threshold, computed as the average of all the sentence scores, to guarantee that only relevant sentences are chosen. All the selected sentences are juxtaposed to compose the final extract.

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 for the complete data collection made available by CLEF, i.e., the sentence in the whole text collection with the best correlation with the question (through the cosine measure). 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 (Ratnaparkhi, 1996);
- 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.

CLEF made available 200 questions, which included the so called “factoid”, “definition”, “temporal” and “list” questions. As defined by CLEF, factoid questions are fact-based questions, asking, for instance, for the name of a person, a location, the extent of something or the day on which something happened; definition questions are questions like "What/Who is X?”; temporal questions ask for facts/events that happened in a determined date or time period; 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 CLEF evaluation: 139 factoid questions, 47 definition questions, 2 temporal 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, the Mean Reciprocal Rank Score 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 for all evaluation measures.

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

Accuracy			Confidence Weighted Score	Mean Reciprocal Rank Score	K1
Questions Judgments	Factoid, definition and temporal	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			Confidence Weighted Score	Mean Reciprocal Rank Score	K1
Question Judgments	Factoid, definition and temporal	List	0.00020	0.0160	-0.5134
Right	3	0			
Wrong	179	10			
Unsupported	4	0			
Inexact	2	2			
Unassessed	0	0			

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.

We attribute the bad results for the following main reasons:

- the sentence level analyses 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;
- the filter we developed was too naïve in trying to classify the questions in only 4 types (“who”, “where”, “when” and “how many” questions); many more variations were used in CLEF.

After CLEF experiments, we believe that simple summarization systems are not enough for question answering tasks, even if these systems are good in what they do. Much more work is needed to bridge this gap.

Acknowledgements

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

References

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, N. 2, pp. 159-165.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken* – PROPOR (Lecture Notes in Artificial Intelligence 2721), pp. 210-218. Faro, Portugal. June 26-27.

Pardo, T.A.S. (2006). *SENTER: Um Segmentador Sentencial Automático para o Português do Brasil*. Technical Report. NILC-TR-06-01. São Carlos-SP, January, 6p.

Ratnaparkhi, A. (1996). A Maximum Entropy Part-of-Speech Tagger. In the *Proceedings of the 1st Empirical Methods in Natural Language Processing Conference*. Philadelphia.

Salton, G. (1989). *Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley.