UNED at WebCLEF 2008: Applying High Restrictive Summarization, Low Restrictive Information Retrieval and Multilingual Techniques

Enrique Amigó, Juan Martinez-Romo, Lourdes Araujo, and Víctor Peinado

NLP & IR Group at UNED, ETSI Informática UNED
c/ Juan del Rosal, 16. E-28040 Madrid, Spain
{enrique, juaner, lurdes, victor}@lsi.uned.es

Abstract. This paper describes our participation in the WebCLEF 2008 task, targeted at snippet retrieval from new data. Our system assumes that the task can be tackled as a summarization problem and that the document retrieval and multilinguism treatment steps can be ignored. Our approach assumes also that the redundancy of information in the Web allows the system to be very restrictive when picking information pieces. Our evaluation results suggest that, while the first assumption is feasible, the second one is not always true.

1 Introduction

The WebCLEF 2008 task has been defined in a similar way to the previous edition. Systems are asked to return a ranked list of snippets extracted from the 1000 web documents identified using the Google web search engine. Multiple languages are covered by the queries and retrieved documents. This task inherits several aspect from Information Retrieval, Summarization and Question Answering tasks. Our approach, as we will describe, is oriented to summarization strategies.

2 Assumptions

Participants are provided with a topic title, a description of the information need, the languages in which the information must be returned, a set of known sources, and a set of queries and their relevant web pages retrieved using Google. The snippets returned by the system must cover the information need without introducing any redundant information already included in the known sources or in other retrieved snippets. Our approach makes the some assumptions that will be tested in the following sections, namely:

1. The terms included in the queries are unambiguous. For instance, "machine translation" (topic 41) refers to systems that translate text from one language into another.

2. Snippets written in different languages tend to contain non redundant information. This assumption avoids the management of multilingual texts that would require additional processing time and linguistic resources.
3. It is possible to find enough information in the Web to build a report containing only sentences that satisfy all requirements established by all basic summarization techniques.
4. The information needs described in topics correspond to the most frequent information returned by Google. This assumption is feasible when queries are defined manually in order to obtain a relatively clean initial ranking.

3 System Architecture

Our system has been implemented over the system described in [1]. In our approach the set of candidate snippets are re-ranked as they are added to the solution. The considered features are:

Noise elimination. Sentences containing more than 5% of words with non-alphabetical characters are discarded. This step removes noisy snippets from the sources.

Snippet length. Sentences containing less that 50 words or more than 200 bytes are removed.[1]

Query terms. The system awards snippets containing query terms, specially when they appear at the beginning of the snippets.[2]

Document relevance. We consider the relevance of the document from which the snippet has been extracted. Initially, we model the document relevance counting the number of query terms appearing in the document.

Centrality. We compute the *vector similarity* described in [1] between the candidate snippet and the rest of candidates. The centrality is the averaged similarity to all candidates.

Redundancy. In a first step, as in [1], we do not consider snippets exceeding a certain similarity threshold with respect to any other snippet in the known sources. In addition, a quantitative redundancy measure is computed by considering the maximum similarity with respect to previously picked snippets.

Key terms contribution. [2] showed that the distribution of key terms has a relevant role in Information Synthesis tasks. Following the approach described in [3], for each topic, we have produced a list of 100 key terms by considering words located immediately before a verb. In order to cover all languages without requiring linguistic processing or big lexicons, we have considered just auxiliary or common verbs such as "is" or "has". After testing several configurations, we have included in the list only those key terms that appear before a verb more than 10 times in the document ranking and in more than 10% of the cases.

[1] Our exploratory studies showed that a minimum length reduces the number of non informative snippets and the maximum length awards the recall of different contents.

[2] The exploratory tentatives have suggested that snippets containing a query term at the beginning are usually more focused on it.

Finally, we have consider the number of key terms appearing in the sentences that didn't appear in previously selected snippets.

In order to compute the snippet score, we calculate the harmonic mean (Rijsbergen's F measure) over Centrality, Redundancy, Key Term Contribution, Document Relevance and Query Terms. Each feature is previously normalized for all the candidate snippets. The motivation for using the harmonic mean rather than other combining criterion is that it is very sensitive to decreases in any of the features, because we expect to find snippets satisfying all requirements at the same time.

3.1 First Variant: More Sophisticated Document Retrieval Step

In order to to test the validity of the assumptions described in Section 2, we have applied information retrieval techniques to select a subset of documents from which our system extracts the snippets. The idea is to select the more relevant documents with respect to several queries composed of terms obtained from different sources. These sources depend on the languages in which the topic is described. We construct an *extended query* for each language in which the system provides some query for the topic. There is always an English extended query composed of terms extracted from the English title and description. This extended query is expanded with terms obtained from the English queries of the topic, if there are any.

For other languages, we translate the topic description from English to the corresponding language[3] and we extract the query terms from this translation and from the queries provided in the considered language. The document relevance is computed following the traditional vector space model which computes the relevance as the minimum cosine distance. The proposed document selection has been implemented using Lucene. For each query we only take the first 50 retrieved documents, which are expected to be more relevant.

3.2 Second Variant: Eliminating Cross-Lingual Redundancy

The second variant consists of a slight modification of our original proposal: we have included a filter in order to eliminate cross-lingual redundancy over pairs of snippets. This filter also uses Google's translation tools and proceeds as follows: 1) automatically detect the language of both input snippets; 2) when necessary, translate each snippet into English (we use machine-translated English as a kind of interlingua to easily compare snippets); 3) remove stop words and; 4) compute words overlap between the two resulting snippets. If the overlap between a candidate snippet and any previously added snippet exceeds a given threshold, it is discarded.

[3] We used Google's Language API services. See `http://code.google.com/apis/ajaxlanguage/documentation` for further details.

Table 1. Results

System	Character Precision	Character Recall
Original approach	0.22	0.21
First variant	0.18	0.17
Second variant	0.21	0.20

4 Conclusions

Analyzing the failures across topics, we have seen that: 1) there are not ambiguous query terms that could affect the results; 2) avoiding redundant information among snippets written in different languages does not contribute to the results; 3) for all topics, the system has found snippets that satisfy all summarization restrictions. These observations suggest that our first three assumptions are correct.

However, the analysis of results across topics suggests that assuming the most frequent information in documents is correlated with the information needs is not applicable to this corpus. In fact not all queries are designed to produce a clean initial set of relevant documents for a given information need. On one hand, some information needs are scattered in two queries: e.g. the "Vanellus" and "collect lapwing eggs" are two independent queries launched to Google to generate the initial ranking for topic 1, and the relevant documents are actually associated to both queries simultaneously. On the other hand, the initial Google queries are not sufficiently precise producing non-relevant documents. For instance, "Algorithms, Data structures and Complexity" do not appear in the general query "computer algorithms contest". In addition, in some cases a bag of words is not enough for detecting the information need, as in "causes of the schizophrenia" vs. "schizophrenia causes".

Our first system tried to solve these problems by including some additional information retrieval techniques but obtained worse results. It seems that it is necessary to analyze more deeply the information need by applying e.g. Question Answering techniques to match the snippet content with the information needs, and to make use of more sophisticated Information Retrieval techniques to tackle the deficiencies of the user query.

References

1. Jijkoun, V., de Rijke, M.: Using Centrality to Rank Web Snippets. Lecture Notes in Computer Science 5152. 737–741. Springer-Verlag. 2007.
2. Amigó, E., Gonzalo J., Peinado V., Peñas, A., Verdejo, F.: An empirical study of information synthesis tasks. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL04). 2004.
3. Amigó, E., Gonzalo J., Peinado V., Peñas, A., Verdejo, F.: Using syntactic information to extract relevant terms for multi-document summarization. Proceedings of the 20th international conference on Computational Linguistics (COLING04). 2004.