Dublin City University at WebCLEF 2007

Sisay Fissaha Adafre
School of Computing, DCU
sfissaha@computing.dcu.ie

Abstract

This paper describes our participation in the Multilingual Web Track (WebCLEF) 2007.

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; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Web retrieval, Questions beyond factoids, Importance ranking, Duplicate removal

1 Introduction

WebCLEF 2007 deals with identifying relevant bits of information from online sources that meets the information needs of an expert user writing an article on a topic. Such information need is expressed in terms of short title of the topic, description of the goals and the intended audience of the article, known online sources that the user considers relevant for the topic and a set of google queries. Given such input, systems are expected to return a ranked list of relevant snippets from the collection provided for the task. The collection consists of the top 1000 (at most) hits from Google for each of the retrieval queries specified in the topic, or for the topic title if the queries are not specified. For this task, we devised a simple aggregative ranking approach which combines evidences from the different elements of the topic description which we will describe in the next sections.

2 System Description

Our approach consists of the following two steps:

- Preprocessing of the collection
- Ranking of the snippets

Preprocessing of the snippets basically consist of splitting the documents in the collection into smaller snippets where each constitutes an information nugget. We consider sentences as our basic information nugget and we split the documents into the constituting sentences or snippets. Following this, we rank the snippets based on their importance to the topic.

3 Ranking Snippets

Snippets are ranked based on their similarity with the topic description. As mentioned in Section 1, the topic is described by multiple informational items, i.e. topic title, google queries, and known online sources. We devised three methods for ranking snippets using these topic descriptions, i.e. baseline, filtering, and parsing.

Baseline

This approach involves a set of preprocessing steps to the topic descriptions. First, the documents from the known online sources are split into sentences or smaller snippets. We then remove stopwords from these sentences. We also remove stopwords from topic titles. The revised topic description consists of the content words from the topic titles, google queries, and a set of snippets (from known-online-source) each represented by its content words (important snippets).

For each candidate snippet, we compute its rank as follows. We first remove stopwords from the candidate snippet. Then we compute its word overlap similarity with the topic titles, google queries and each of the snippets from the known online sources. The final score for the snippet is the average score of these similarity scores.

Filtering

In this approach, we preprocess the topic descriptions as shown in the *Baseline*. Unlike the *Baseline*, this method applies a two-step ranking process.

First, we combined the content words of the topic titles and google queries, and formed a set of query terms. We computed the word overlap score between this set and the candidate snippets represented by their content words. We took candidate snippets that fall above a certain threshold.

Following this, we reranked the resulting candidate snippets based on their average word overlap score with a set of snippets (from known-online-source) each represented by its content words.

Parsing

The above two methods are based on a relatively language independent (and shallow) methods that can easily be applied to a problem with multilingual requirement. However, the underlying scoring method uses simple word overlap metric, where the topic description and the candidate snippets are represented by their content words. The content words are selected if they do not belong to a precompiled stopword list.

In this approach, we want to impose more constraint on the choice of content words. We considered only the top 20 ranking webdocuments for processing. Like the *Filtering* method, we filtered the candidate snippets based on their word overlap with the topic title and google queries.

The resulting list of candidate snippets are reranked as follows. We parsed each of the candidate snippets and the set of snippets from the known online sources using a Lexical Functional Grammar based parser [1]. Each of the candidate snippets is represented by the corresponding set of head words we obtain from the parse tree. We then compute a word overlap score for each candidate snippet with the snippets from known online sources. The final score for a candidate snippet is the average of these scores. For non-English topics, the system applies the baseline method of computing the rank, i.e. without applying parsing.

Finally, a simple duplicate removal procedure is applied to the outputs of the above three systems. We sort the output of the above systems in descending order of their scores. For each high ranking snippet, we compute its wordoverlap scores with other snippets that are ranked lower and remove those with a word overlap score above a threshold value. We then return the top 30 snippets as our final result set.

4 Result

Table1 presents the results of our three runs. The precision and recall measures in the result table are defined as follows [2].

- Nugget-based (resp., character-based) Recall: the number of the all identified nuggets (resp., their character length) which are covered by the snippets of a system S, divided by the total number of nuggets (resp., their total character length)
- Precision: the number of characters that belong to at least one span linked to a nugget, divided by the total character length of the system's response.

	Precision	Recall
Simple	0.078	0.050
Filtering	0.076	0.046
Parsing	0.103	0.071

Table 1: Results.

Overall the scores are very low compared to the top scoring system (Precision: 0.202 and Recall: 0.256). Parsing seems to improve the result to some extent. Since the improvement is small, further experiment needs to done to see the actual contribution of applying parsing to the overall improvement in the scores. On the other hand, the first two approaches did not show any significant differences. We observed that our system returns mostly short snippets (compared to other systems) since we split the documents into sentences. In some cases, the splitting process returns fragmented sentences containing incomplete information. Larger textual units, such as paragraphs, may be an appropriate retrieval unit for the current task.

5 Conclusion

We applied three different methods for the task of identifying important snippets from the Web. Our aim is to devise a simple and generic method that can be applied in the context of Multilingual Web retrieval. We also applied deeper natural language analysis method which is mainly targeted for English language. The experimental results show that there is some room for improvement. In the future, we will investigate the contribution of redundancy information and structural properties of the Web documents for ranking the snippets.

References

[1] Aoife Cahill, Michael Burke, Martin Forst, Ruth Odonovan, Christian Rohrer, Josef Genabith, and Andy Way. Treebank-based acquisition of multilingual unification grammar resources. *Research on Language and Computation*, 3(2):247–279, July 2005.

[2] WebCLEF, 2007. WebCLEF: The CLEF Crosslingual Web Track `http://ilps.science.uva.nl/WebCLEF/WebCLEF2007`.