Link-Based vs. Content-Based Retrieval for Question Answering Using Wikipedia

Sisay Fissaha Adafre, Valentin Jijkoun, and Maarten de Rijke

ISLA, University of Amsterdam
{sfissaha,jijkoun,mdr}@science.uva.nl

Abstract. We describe our participation in the WiQA 2006 pilot on question answering using Wikipedia, with a focus on comparing link-based vs content-based retrieval. Our system currently works for Dutch and English.

1 Introduction

The WiQA (Question Answering using Wikipedia) task derives much of its motivation from the observation that, in the Wikipedia context, the distinction between reader and author has become blurred; see [3] for an overview of the task. Given a topic (and associated article) in one language, identify relevant snippets on the topic from other articles in the same language or even in other languages. In the context of Wikipedia, having a system that offers this type of functionality capability is important for readers as well as authors.

In this report, we describe our participation in the WiQA 2006 pilot. We took part both in the monolingual and bilingual tasks. Our approach consists of three steps: retrieving candidate snippets, reranking the snippets, and removing duplicates. We compare two approaches for identifying candidate snippets, i.e., link-based retrieval and traditional content-based retrieval. Section 2 presents an overview of the system; Section 3 describes our runs; Section 4 provides our results, and Section 5 presents conclusions.

2 System Description

We briefly describe our mono- and bilingual systems for the WiQA 2006 pilot. The main components of the monolingual system are aimed at addressing the following subtasks: identifying relevant snippets, estimating sentence importance, and removing redundant snippets. We briefly discuss each of these in turn.

We apply two approaches to identifying relevant sentences. The first exploits the peculiar characteristics of Wikipedia, i.e., the dense link structure, to identify relevant snippets. The link structure, and in particular, the structure of the incoming links (similar to Wikipedia's "What links here" feature), provides a simple mechanism for identifying relevant articles. If an article contains a citation to the topic then it is likely to contain relevant bits of information about the topic

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 537–540, 2007.

since the hyperlinks in Wikipedia are part of the content of the article. Since hyperlinks are created manually by humans, this approach tends to produce little noise. However, not all mentions of a topic may be hyperlinked which may cause some recall problems. Furthermore, coreferences may also need to be resolved in order to improve recall. Our second approach to identifying relevant sentences is based on an out-of-the-box Lucene [4] index of Wikipedia articles. From retrieved articles, we extract sentences that mention (sufficiently large parts of) the topic for which we are attempting to locate relevant information. We expect the latter content-based approach to be more recall-oriented than the link-based approach, potentially picking up more noise in the process.

For estimating sentence importance, we combine several types of evidence: retrieval score, position in the article, and a graph-based ranking score. We normalize the retrieval scores (obtained in the first step) to values between 0 and 1 by dividing each retrieval score with the maximum value. Furthermore, the earlier a sentence appears in an article, the more important we assume it to be. Sentence positions are converted into values between 0 and 1 in which the sentence at position 1 receives the maximum graph-based score (explained below), and subsequent sentences receive a score which is a fraction of the maximum graph-based score using the formula proposed in [5]. Finally, graph-based scoring allows us to rank sentences by taking into account the more global context of sentences, which consists of a representative sample of articles that belong to the same categories as the topic. The resulting corpus serves as our importance model by which we assign each sentence a score; see [2].

For the final filtering step, we compare each candidate sentence with the sentences in the main article, and sentences that appear before it in the list of sentences ranked by importance score. We used simple word overlap for measuring sentence similarity. We keep the maximum similarity score and subtract it from the sentence score. We sort the list again in decreasing order of the resulting scores, and return the top 10 sentences as output.

As to our approach to the multilingual (Dutch-English) task, given a topic in one language, we need to find important snippets in other Wikipedia articles of the same language or other languages. The main challenge is to ensure relevance and novelty across languages. To measure similarity among snippets from different languages we used a bilingual dictionary generated from Wikipedia itself (from the titles of corresponding articles in the two languages) and applied the method to Dutch and English articles. Briefly, given a topic in one language, we identify important snippets in the language of the topic translate the topic into other languages (using the bilingual dictionary), identify important snippets in the translated topics, and filter the resulting multilingual list for redundant information, again using the bilingual dictionary; see [1].

3 Runs

We submitted a total of seven runs: six monolingual runs for Dutch and English (three runs per language), and one is a Dutch-English bilingual run. All the

monolingual runs use the approach outlined in Section 2. Except for the use of stopword lists, the method is completely generic and can be ported to any language with relative ease. The runs differ in the methods adopted for acquiring the initial set of relevant sentences: link-based (*Link*), content-based (*Ret*), or a combination of the two (*LinkRet*).

For the bilingual run we consider Dutch and English articles. The source topics can be in any of these languages. As mentioned in the previous section, we build on top of our monolingual approach. We used the output of the third monolingual run as an input for the bilingual filtering.

4 Results

We present the results of our seven runs. The results are assessed based on the following attributes: "supported," "importance," "novelty," and "non-repetition." The summary statistics that we report are *Avg. Yield* (the total number of supported, important, novel, non-repeated snippets in the top 10, per topic, averaged over all topics); MRR (mean reciprocal rank of the first supported, important, novel, non-repeated snippet in the top 10; p@10 (the fraction of supported, important, novel, non-repeated snippets in the top 10 snippets, per topic, averaged over all topics). See the overview paper for the WiQA pilot task for further details [3]. Table 1 shows the results of the seven runs.

The scores for the *Link* based monolingual runs are the best, suggesting that link-based retrieval provides a more accurate initial set of relevant sentences on which the performance of the whole system largely depends. The content-based approach seems to introduce more noise as shown by the scores of the *Ret* and *LinkRet* based monolingual runs for both languages. Contrary to our expectation, the combination of the two methods performed worse for English.

In order to see whether our approaches return similar sets of snippets, we compared the result sets, i.e. sets of supported, important, novel and non-repeated snippets returned by the methods. While *Link* returned 353 *important* snippets and *Ret* returned 325 snippets in top 20, the joint pool of the two methods contains 484 important snippets. This indicates that there is a potential room for improvement in combining the link-based and retrieval-based methods for importance estimation. Unfortunately, our combination method used in the run *LinkRet* did not perform well. In future work we will focus on other ways to effectivly combine different sources of evidence for this task.

Table 1. Results for the English and Dutch monolingual tasks and the Dutch-English bilingual task

	English			Dutch			English-Dutch		
	Avg. Yield	MRR	p@10	Avg. Yield	MRR	p@10	Avg. Yield	MRR	p@10
Ret	2.938	0.523	0.329	3.200	0.459	0.427			
Link	3.385	0.579	0.358	3.800	0.532	0.501			
LinkRet	2.892	0.516	0.330	3.500	0.532	0.494	5.03	0.518	0.535

The scores for the bilingual run tend to be higher than our monolingual scores. This may be due to the fact that most of the topics are Dutch topics that are mostly short and additional snippets are likely to be new. Furthermore, the snippets can come from both Dutch and English encyclopedias which also contributes to finding good snippets.

5 Conclusion

We have described our participation in the WiQA 2006 pilot task. Our approach consists of three steps: retrieving candidate snippets, reranking the snippets, and removing duplicates. We compared two approaches for identifying candidate snippets: link-based retrieval and traditional content-based retrieval. The results show that the link-based approach performed better than the content-based approach, and although a combined approach showed a relatively poor performance, our analysis suggests potential room for improvement in combining the two methods.

Acknowledgments

Sisay Fissaha Adafre was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 220-80-001. Valentin Jijkoun was supported by NWO under project numbers 220-80-001, 600.-065.-120 and 612.000.106. Maarten de Rijke was supported by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.-065.-120, 612-13-001, 612.-000.106, 612.066.302, 612.069.006, 640.001.501, 640.002.501, and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

References

[1] Fissaha Adafre, S., de Rijke, M.: Finding similar sentences across multiple languages in wikipedia. In: EACL 2006 Workshop on New Text, Wikis and Blogs and Other Dynamic Text Sources (2006)
[2] Fissaha Adafre, S., de Rijke, M.: Learning to identify important biographical facts (Submitted)
[3] Jijkoun, V., de Rijke, M.: Overview of WiQA 2006 (In This volume) (2006)
[4] Lucene: The Lucene search engine, `http://lucene.apache.org/`
[5] Radev, D.R., Jing, H., Sty, M., Tam, D.: Centroid-based summarization of multiple documents. Information Processing and Management 40(6), 919–938 (2004)