The University of Amsterdam at WiQA 2006

Sisay Fissaha Adafre Valentin Jijkoun Maarten de Rijke
ISLA, University of Amsterdam
sfissaha,jijkoun,mdr@science.uva.nl

Abstract

This paper describes our participation in the WiQA 2006 pilot on question answering using Wikipedia. We present an analysis of the results of our system for both monolingual and bilingual runs. The system currently works for Dutch and English.

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

Question answering, Questions beyond factoids, Importance ranking, Novelty detection, Duplicate removal

1 Introduction

The WiQA 2006 pilot deals with access to the content of Wikipedia, the free online encyclopedia. At WiQA, information access is considered both from a reader's point of view and from an author's point of view—WiQA 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. The overlap in the roles of the different types of users in Wikipedia motivates new modes of information access, ones that can support the emerging dual roles identified above.

WiQA 2006 attempts to address this issue by formulating the task definition as follows. 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 both for using Wikipedia as a reader and as an author. It provides effective access to additional relevant information in Wikipedia that is not found in the main article on the topic, which, we believe, is important for both use cases.

In this report, we describe our participation in the WiQA 2006 pilot. We took part both in the monolingual and bilingual tasks. Section 2 presents an overview of the system. Section 3 describes the runs that we submitted. Section 1 provides an analysis of the results. Finally, Section 5 presents some concluding remarks.

2 System Description

We briefly describe our mono- and bilingual systems for the WiQA 2006 pilot; for a far more detailed description of our monolingual system, please consult [2]

2.1 Monolingual Task

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 now discuss each of these in turn.

2.1.1 Identifying relevant snippets

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 second approach makes use of standard document retrieval techniques. We briefly review the two approaches below.

Link based retrieval. The link structure, particularly, 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 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, due to inconsistencies in manual processing and also editing requirements, 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.

We devised a simple coreference resolution method for a particular class of coreferences, i.e., last name resolution for a person. This method checks whether a person name appears at least once in the article as hyperlink. Then the person's name will be tokenized into words and the last word will be taken as last name of the person. The last name will then be replaced by the full name. This ommits a large class of coreferences such as pronouns and definite descriptions. However, the use of pronouns in Wikipedia is relatively low. Furthermore, current coreference resolution techniques mostly use deep linguistic analysis which is hard to scale up to corpora of the size of Wikipedia.

Lucene based retrieval. We indexed the Wikipedia articles using the Lucene retrieval engine [4]. We assumed each article to constitute a single document in the index. The resulting index is used to retrieve articles that contain information about a topic. We used the title of the topic, which corresponds to the title of a Wikipedia article, as a query to retrieve articles. Since the aim is to identify smaller snippets about the topic, we split articles into sentences and keep only those sentences that contain an occurrence of the topic at hand. For non-person topics, we require the snippets to contain all the tokens of the topic. On the other hand, for person topics, we also consider snippets that contain one of the tokens in the name. Unlike the link based retrieval approach outlined above, which requires strict hyperlink relation, this method only checks whether a snippet contains the title of the topic which may not necessarly imply a hyperlink relation. As a result, the method tends to be recall oriented, identifying more relevant snippets. However, the relaxed criteria also means that the method is likely to pick up more noise.

2.1.2 Estimating Sentence Importance

Once we have an initial set of relevant sentences, the next phase of the processing step is to rank the sentences based on their importance to the topic. For this, we combine several types of evidence. These include retrieval scores, position in the article, and graph-based ranking scores.

Retrieval scores. The topic is used as a query to retrieve the initial set of relevant articles. Articles containing multiple occurrences of the topic will be ranked higher. The assumption is that such higher ranking articles are likely to contain larger number of relevant sentences, which in turn increases the likelihood of getting important and novel sentences. This provides for a relatively weak source of evidence of importance of a sentence in an article. We convert the retrieval scores to values between 0 and 1 by dividing each retrieval score with the maximum value.

Position in article. The position of a sentence in an article is used as an extra indication for its importance for the topic of the article from which the sentence is extracted. The earlier the sentence appears in the article, the more important it is assumed to be. The sentence positions are converted into a value 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].

Graph-based scoring The previous methods make use of local context in assigning relative importance to sentences. Graph-based scoring allows us to rank sentences by taking into account a more global context of the sentences. This global context 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. Once we have our representative corpus, we rank sentences based on their centrality with respect to this corpus. For more detailed discussion of the method, see [2].

The overall score of a sentence is the sum of the above scores.

2.1.3 Filtering

After we compute the scores for the snippets, the next step involves computing a redundancy penalty [5]. For this, we sort the snippets by decreasing order of their scores. We compare each candidate sentence with the sentences in the main article, and sentences that appear before it in the ranked list. We used simple word overlap for measuring sentence similarity. We keep the maximum similarity score and subtract it from the sentence score. In all our similarity computations, we remove stopwords. We sort the list again in decreasing order of the resulting scores. The top 10 sentences constitutes the snippets in our submitted runs.

2.2 Multilingual Task

The multilingual task extends the problem of finding important snippets in one language to multiple languages in Wikipedia: given a topic in one language, find important snippets in other Wikipedia articles of the same language or other languages. The major challenge in this task concerns ensuring relevance and novelity acrose multiple languages. This in turn means that we need to have some mechanism of measuring similarity among snippets from different languages. This may be achieved using different approaches, i.e., using a biligual dictionary or machine translation system. For this task, we only used a bilingual dictionary generated from Wikipedia itself. We applied the method on Dutch-English language pairs.

2.2.1 Steps

Identifying important snippets in a multilingual task consists of several monolingual extraction plus multilingual filtering. Specifically, given a topic in one language, we

- identify important snippets in the language of the topic (primary language).
- translate the topic into other languages
- identify important snippets in the translated topics (secondary languages)
- filter the resulting multilingual list for redundant information

The ranked list consists of three types of snippets. The first set contains the snippets extracted from the primary language. These snippets are extracted using the monolingual procedure presented in the Section 2.1 (primary snippets). The second set contains snippets coming from the main article in the secondary language (secondary main article snippets). These are automatically considered as relevant and are included in the main list. The third type of snippets are extracted from other articles in secondary language (secondary article snippets), again using the monolingual algorithm, but this time on the secondary language. The snippets are ordered such that the primary snippets comes at the top followed by secondary main article snippets which are then followed by the secondary other article snippets.

The above ranked list and the main article in the primary language are input to the filtering module, which goes through the list and removes duplicates. In the next section we briefly summarize the method we adopt for multilingual similarity which forms the core of the filtering module.

2.2.2 Using a Bilingual Lexicon

We now describe the method we used for identifying similar text across different languages. It relies on a bilingual lexicon which is generated from Wikipedia using the link structure. The bilingual lexicon consists of the Wikipedia page titles that typically represents concepts or entities that have entries in Wikipedia. Therefore, similarity measure is based on concept or article title overlap. We used page title translations as our primitives (as main features) for the computation of multilingual similarity. For each Wikipedia page in one language, translations of the title in other languages, for which there are separate entries, are given. This information can be used to generate a bilingual translation lexicon. Most of these titles are noun phrases and are very useful in multilingual similarity computations. Most of these noun phrases are already disambuiguated. They are mostly content bearing terms and may consist of single word or multiword units.

We used the Wikipedia redirect feature to identify synonymous expression. In Wikipedia, the redirect feature is used to map several titles into a canonical form. For a more detailed description see [1].

3 Runs

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

3.1 Monolingual Runs

All the monolingual runs use the approach outlined in Section 2. Observe that except for the use of stopwords lists, the method is generic and can be ported to a new language with relative ease.

The runs differ in the methods adopted for acquiring the initial set of relevant sentences. As mentioned in Section 2.1.1, we identified two ways of identifying relevant sentences. The first run uses the link based approach whereas the second uses retrieval based approach. The third run combines the retrieval based approach with link based filtering in identifying relevant sentences. The latter two runs use the retrieval score in computing the overall score of sentence importance.

3.2 Bilingual Run

The bilingual run considers Dutch and English language pairs. The source topics can be in any of these languages. As mentioned in Section 2.2, the method is built on top of our monolingual approach. We used the output of the third monoligual run as an input for the bilingual filtering. The bilingual run makes use of the bilingual lexicon generated automatically from Wikipedia for identifying duplicates across different languages.

4 Results

We present the results of our seven runs. The results are assessed based on the following attributes: "supported", "important", "novel" and "not repeated". The summary statistics used as evaluation measures are:

- Yield at top 10 snippets: total number of supported important novel non-repeated snippets for all topics among the top 10 snippets (*yield*);
- Average yield per topic (top 10 snippets): previous, devided by the number of topics (*Avg. Yield*) ;
- MRR (top 10 snippets): Mean Reciprocal Rank of the first supported important novel non-repeated snippet among top 10, according to the system's ranking(*MRR*);
- Precision at top 10 snippets: the number of supported important novel non-repeated snippets among top 10 snippets per topic, divided by the total number of top 10 snippets per topic (*Precision*).

The test files for the English and Dutch monolingual tasks consist of 65 and 60 topics respectively. The corresponding test file for bilingual task consists of 60 topics.

Table 1 shows the results of the seven runs. As mentioned in Section 3, our monolingual runs differ on how the initial set of relevent sentences are acquired. In Table 1, the three different approaches are indicated by *Ret* (retrieval only approach), *Link* (link only based approach), and *LinkRet* (the combination of the two methods). The columns in Table 1 are as described above.

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

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

Overall the scores for *link* based monolingual runs are the best. This shows that link based retrieval approach provide a more accurate initial set of relevant sentences on which the performance of the whole system largely depends. The retrieval 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 the three approaches return similar sets of snippets, we counted the number of snippets that are found in the result sets of the three methods for English monolingual runs. The three methods have 305 snippets in common in their result sets. Of these snippets, 103 snippets are judged good snippets where the number of good snippets returned by the three methods are; 220-*Link*, 191-*ret*, and 188-*linkret*.

Furthermore, the three methods return no good snippets for the following topics, *Brooks Williams*, *Wing Commander (film)*, *Christian County, Illinois*, *White nationalism*, *Telenovela*

database, *Oxygen depletion*. Although the potential cause for the poor performance on these topics may vary acrose the topics, a brief look at the outputs for these topics reveals some possible sources of problems. For some of these topics, the retrieval component returned very few candidate snippets, e.g. *Brooks Williams* and *Oxygen depletion*. For others, it returned a large number of similar snippets, e.g. towns and cities for *Christian County, Illinois* and different entries of *Telenovela database*, which are judged irrelevant or redundant by the assessors. Some topics have ambiguous titles, e.g. *Wing Commander (film)*, as indicated by its result set which contains snippets from articles with similar titles, e.g *Wing Commander (computer game)*.

On the other hand, the three methods return more than 5 good snippets for the following topics: *Center for American Progress*, *Atyrau*, *Saitama Prefecture*, *Kang Youwei*, *Philips Records*. Further examination of the outputs of the three methods for these topics shows that the methods tend to return similar sets of good snippets. Overall most of the scores for our English monolingual runs are above the median score. The scores for our best run is close to the maximum score.

A similar analysis on the output of the Dutch monolingual runs shows similar patterns. Some topics are very hard for all methods. For example, all methods returned no snippets at all for the following topics; *Vclav Havel*, *MG (auto)*, *Tsjetsjeni*, *Caro*, *Socit Nationale des Chemins de fer Franais*, *TVR (auto)*. This is mainly due to the special characters in the titles of the topics which we did not anticipate and made it impossible to identify candidate snippets. On the other hand, all methods performed well on the following topics; *Gilera*, *Columbia-universiteit*, *Albert Einstein*, *Zwarte rat*, *NSU*, *Slangendrager* .

The scores for the billingual run is based on the output of the *LinkRet* retrieval component. The scores tend to be higher than the 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 new. Furthermore, the snippets can come from both Dutch and English encyclopedias which also contributes to finding good snippets.

5 Conclusion

This paper described our participation in the WiQA 2006 pilot. Our approaches consist basically of the following three steps: retrieving candidate snippets, reranking the snippets, and removing duplicates. We compared two approaches for identifying candidate snippets, i.e. link based retrieval and the traditional document based retrieval. The result showed that the link based approach performed better than the document retrieval based approach. The results showed that overall our system performed well. However, there is a lot of room for improvements. In the future, we want to compare different reranking methods and also perform detailed error analysis of the output of our system.

6 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, and 640.002.501.

References

[1] Sisay Fissaha Adafre and Maarten de Rijke. 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] Sisay Fissaha Adafre and Maarten de Rijke. Learning to identify important biographical facts. In *Submitted*, 2006.

[3] Valentin Jijkoun and Maarten de Rijke. Overview of WiQA 2006. In *This volume*, 2006.

[4] Lucene. The Lucene search engine. `http://lucene.apache.org/`.

[5] Dragomir R. Radev, Hongyan Jing, Magorzata Sty, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938, 2004.