University of Alicante at WiQA 2006

Antonio Toral Ruiz, Georgiana Puşcaşu*, Lorenza Moreno Monteagudo
Rubén Izquierdo Beviá, Estela Saquete Boró
Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
{atoral,georgie,loren,ruben,stela}@dlsi.ua.es

Abstract

This paper presents the participation of University of Alicante at the WiQA pilot task organized as part of the CLEF 2006 campaign. For a given set of topics, this task presupposes the discovery of important novel information distributed across different Wikipedia entries. The approach we adopted for solving this task uses Information Retrieval, query expansion by feedback, relevance and novelty re-ranking, as well as temporal ordering. Our system has participated both in the Spanish and English monolingual tasks. For each of the two participations the results are promising because, by employing a language independent approach, we obtain scores above the average. Moreover, in the case of Spanish, our result is very close to the best achieved score. Apart from introducing our system, the present paper also provides an in-depth result analysis, and proposes future lines of research, as well as follow-up experiments.

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;

General Terms

Measurement, Performance, Experimentation

Keywords

Information Retrieval

1 Introduction

Wikipedia[1] is a multi-lingual web-based, free content encyclopedia, continuously updated in a collaborative way. It may be seen as a paradigmatic example of a huge[2] and fast-growing source of written natural language.

Several inherent characteristics of this resource, such as its continuous growing nature, its general domain coverage, as well as its multilinguality, make Wikipedia a valuable resource for the Natural Language Processing (NLP) research field. The NLP community has only just lately

* Currently on research leave from University of Wolverhampton, United Kingdom.
[1] www.wikipedia.org
[2] By July 2006, the English version alone contains more than 1,250,000 entries

become aware of this fact and started investing research effort in possible ways of exploiting Wikipedia within strategic areas such as Question Answering [2] or Knowledge Acquisition [8].

WIQA[3] is a pilot task at CLEF 2006[4] exploiting the fact that, in Wikipedia, the distinction between author and reader has become blurred. The aim of the task is to discover how Information Retrieval and NLP techniques can be effectively used to help readers and authors of articles get access to information spread throughout Wikipedia rather than stored locally on a single page [4]. In a nutshell, WiQA is about collecting information about a certain topic not yet present on its page, thus avoiding data sparseness and unifying related content distributed among different entries. The motivation to launch this task lies in the already pointed out challenges that Wikipedia poses to the NLP community.

This paper is organized as follows. The following section presents a description of our approach and the developed system. Section 3 describes the experiments submitted to the WIQA task. Afterwards, in section 4, we present and comment the obtained results. Finally, in section 5, conclusions are drawn and future lines of research are pointed out.

2 System description

Inspired by the Novelty and the QA tasks at TREC, the WiQA pilot task aims at recovering information not explicitly mentioned on a page, but distributed across the entire encyclopedia. The envisaged participating systems should help provide access to, author and edit Wikipedia's content. They should, mainly, be able to locate relevant and new sentences within the Wikipedia document collection, in response to a topic.

The WiQA task could be tackled from two perspectives, by employing either Information Retrieval methods, or Question Answering capabilities. Our approach embraces the IR strategy. Information Retrieval [3] is a Natural Language Processing application that, given a query and a document collection, returns a ranked list of relevant documents in response to the input query. IR usually comprises two stages. The first is a preprocessing phase which is carried out offline and consists of indexing the document collection. Its aim is to represent the documents in a way that makes it easier and more efficient to store and interrogate the collection. The second step, carried out online, consists of the actual retrieval of relevant documents as answer to an input query.

The architecture of our system is depicted in Figure 1. IR forms the core part of the system, being used to retrieve the documents relevant to the most meaningful terms in the topic document. For example, considering the topic *Alice Cooper*, we first extract the most meaningful terms in the supporting topic document. Then, in order to retrieve documents not only mentioning *Alice Cooper*, but also belonging to the domain defined by the extracted terms, we search the collection using these relevant terms.

For the above presented purposes, we have employed a probabilistic open source IR library called Xapian [1]. Besides the probabilistic search capability, in which the most relevant words are given increased weight, it allows boolean searches with operators that affect the query words, thus placing user-defined constraints on the search. These operators allow the user to specify, for example, that the desired terms occur in close proximity to each other. Another useful feature of Xapian is the possibility to receive feedback. By using this technique, Xapian can extract relevant terms for the query and carry out an expansion of it. The system first performs a basic retrieval and then gives the user the opportunity to select a set of documents considered relevant. At the next step, Xapian extracts from the selected documents relevant terms for query expansion. Finally, a second retrieval is performed by adding these terms to the query.

Due to the nature of WiQA pilot task the indexation has been performed at sentence level, that is sentences have been resembled to complete documents, therefore each indexed document is made up of only one sentence. This makes it straightforward to retrieve directly sentences in order to be compliant with the desired output of the system. In consequence, our system comprised

[3] http://ilps.science.uva.nl/WiQA/

[4] http://www.clef-campaign.org/

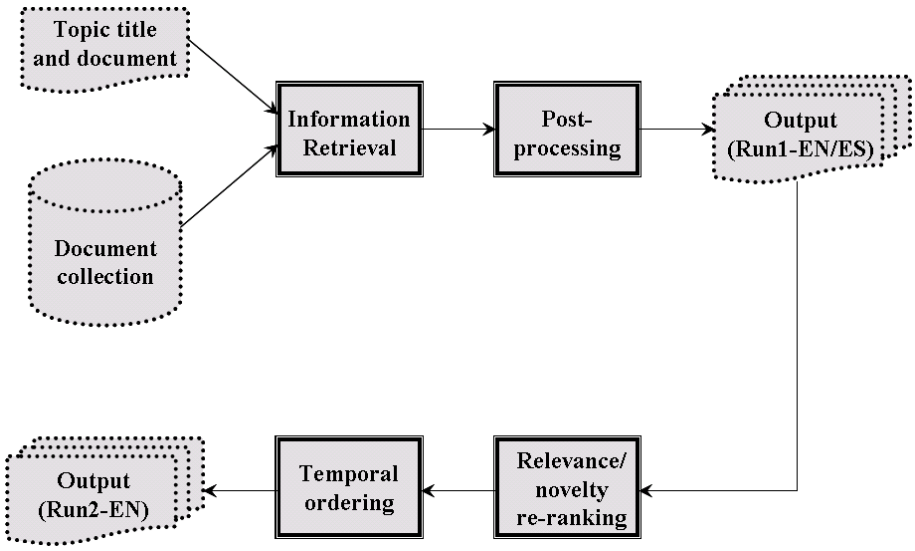

Figure 1: System architecture

a preprocessing phase that consisted of document sentence splitting and SGML tags removal. Finally, all the sentences contained in the document collection are indexed.

At the retrieval stage, Xapian has been configured using options and parameters to be presented in detail in the next section 3. Once the relevant sentences have been identified, they have been passed through a post-processing stage consisting of the following actions:

- Eliminate those sentences which belong to the topic supporting document (the document identifier corresponding to the sentence matches the one of the topic)
- Eliminate those sentences that belong to documents linked from the query document (there is a link in the topic supporting document that points to the document of the sentence)

At this stage, a core set of sentences possibly relevant and important for the topic in question has already been delimited. In the case of the Spanish monolingual task, this set of sentences forms the system output, while, for the English task, they pass through subsequent processing stages, as described in the following.

Our English system proceeds by parsing the set of possibly relevant sentences, as well as the topic document sentences, with the Conexor's FDG Parser [7, 5]. For the two sets of sentences and supporting document titles, the time expressions are also identified and resolved using the temporal expression recogniser and normaliser previously developed by one of the authors [6]. The relevance and novelty of each retrieved sentence with respect to the sentences included in the Wikipedia topic document is then measured, in order to preserve the most relevant sentences to update the content of the Wikipage. Therefore, the sentences manifesting a high degree of similarity with the content of the topic document were characterised by very low scores.

The degree of relevance and novelty of a retrieved sentence with respect to a sentence from the topic's Wikipage was considered to be a weighted measure revealing the percentage of novel named nouns (all uppercase nouns situated in the middle of the sentence), non-matching temporal expressions, novel nouns and verbs included in the former sentence, but not in the latter one.

The retrieved sentences are then ranked with respect to their relevance/novelty score, and passed on to a temporal ordering module. The temporal ordering module labels each retrieved sentence with the first TE occurring in the sentence, or, if no TE is present in the sentence, with the first TE of the title, or, if still no TE is found, with no label. Afterwards, the labelled sentences are interchanged so that their new order reflects their temporal order. The unlabelled sentences thus preserve their rank reflecting their relevance.

3 Description of submitted runs

Our WiQA submission includes one run for the Spanish monolingual task and two runs for the English task. The Spanish run and the first English run have been obtained with the same methodology, but applied to the language specific text collections. The second English run was obtained by employing more sophisticated NLP techniques to rank the set of retrieved sentences according to novelty/relevance, and to order them chronologically.

The first English run, as well as the Spanish run, both employ only the Information Retrieval capabilities of our system. Xapian firstly performs a search for the topic title constrained by the NEAR operator. The NEAR operator helps in locating the topic title words situated in any order within a short distance from each other. The feedback characteristic of Xapian is then employed to extract the important terms from the defined set of relevant documents (we classify as relevant the first 50 retrieved documents). Then the query is expanded with the identified relevant terms and a second retrieval is performed, this time constrained by the PHRASE operator. The PHRASE operator identifies only the sentences where the group of words defining the topic occur together and in the same order as in the query. After post-processing the retrieved sentences by filtering out the ones that occur either in the topic document or in any document linked from the topic document, we preserve only the first 10 resulted sentences and return them as result of the first English run and of the Spanish run respectively.

Our second English run performs, apart from Information Retrieval, a relevance/novelty-based ranking, followed by a temporal ordering stage. Information Retrieval in employed in the same manner and with the same specifications as in the case of the first run. The retrieved sentences together with the topic sentences are parsed with the morpho-syntactic parser and with the temporal expression identifier/normaliser described above. A measure of relevance and novelty is then computed for each retrieved sentence with respect to all sentences from the topic document in turn, and the minimum relevance score obtained will represent its degree of relevance/novelty with respect to the entire topic document. A relevance/novelty ranking of the retrieved sentences is then produced and passed on to a temporal ordering module. The temporal ordering module produces a new order of the sentences that reflects their succession on the temporal axis.

4 Results and discussion

In this section we present and comment on the obtained results. As already stated, we have submitted three runs for WiQA 2006. Two runs represent solutions for the English monolingual task (one using IR only and the other one employing extra re-ranking and temporal ordering capabilities). The third run employs only IR and corresponds to the Spanish monolingual task.

The following two tables summarize the results for the monolingual English (Table 1) and for the monolingual Spanish (Table 2) tasks.

Run ID	**Average Yield**	**MRR**	**Precision**
1	2.98	0.53	0.33
2	2.63	0.52	0.32
MIN	1.52	0.30	0.20
MED	2.46	0.52	0.32
MAX	3.38	0.59	0.37

Table 1: Results for the English monolingual task

For each run, three different measures are provided (average yield, MRR and precision), all measured for the top 10 snippets returned. The average yield represents the average number of supported & novel & non-repetitive & important snippets retrieved. The MRR (Mean Reciprocal

Rank) score refers to the first supported & novel & non-repetitive & important snippet returned. The precision was calculated as the percentage of supported & novel & non-repetitive & important snippets encountered among the submitted snippets. Apart from the results achieved by our runs, the two tables also include the minimum, median and maximum scores obtained for the tasks at hand. We are therefore able to evaluate and compare the performance of our system with respect to other participants.

Run ID	**Average Yield**	**MRR**	**Precision**
1	1.76	0.36	0.22
MIN	1.02	0.29	0.16
MED	1.06	0.30	0.22
MAX	1.82	0.37	0.27

Table 2: Results for the Spanish monolingual task

The results show that our approach using feedback-driven IR has obtained results above the median value, both for English (Table 1) and for Spanish (Table 2). The fact that we score considerably better for English than for Spanish, though using the same approach, (average yield 2.98 EN vs. 1.76 ES, MRR 0.53 EN vs. 0.36 ES and precision 0.33 EN vs. 0.22 ES) might be due to the different size of the Spanish Wikipedia in comparison with the English version. Being the English version notably larger than the Spanish version, there might be many more topic relevant text snippets spread across its entries.

Regarding the run which uses, apart from IR, relevance and novelty re-ranking plus temporal ordering, it has been submitted only for English, as it makes use of language dependent tools. The obtained results have been slightly worse than the ones for the first run. Our expectations were that these post-processing stages would at least bring a slight improvement to the results given by the IR engine alone. However, the performance has decreased. An in-depth analysis is needed to find out the causes of this unexpected behavior. Our opinion is that further investigation is required to improve or discover a more appropriate formula to be employed for measuring the degree of relevance and novelty of a retrieved snippet. Besides, temporal processing should probably be employed at a point when WiQA systems are more mature and the input snippets are more reliable.

5 Conclusions

This paper presents our approach and participation in the WiQA 2006 competition. We propose feedback-driven IR with query expansion in order to retrieve, in response to given Wikipedia entries, relevant information scattered throughout the entire encyclopedia. Moreover, we have introduced a post-processing stage consisting of novelty/relevance ranking and temporal ordering. Our system participated both in the Spanish and English monolingual tasks.

When compared to the other systems presented in this competition, we have obtained good results that situate us above the medium score and quite close to the best result in the case of Spanish. Therefore, we conclude that the proposed approach is appropriate for the WiQA task and we plan to find ways of improving the system's performance.

Several future work directions emerge naturally from a first look and shallow analysis of the results. Firstly, we would like to carry out an in-depth study of the effects induced by applying novelty/relevance ranking and temporal ordering, as the results obtained have not been those expected. Secondly, we aim at furtherly investigating this topic departing from our feedback-driven Information Retrieval approach.

6 Acknowledgements

This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01.

References

[1] Xapian: an Open Source Probabilistic IR library. On line www.xapian.org. Visited 2006-06-01.

[2] D. Ahn, V. Jijkoun, G. Mishne, K. Mller, M. de Rijke, and S. Schlobach. Using wikipedia at the trec qa track. In *The University of Amsterdam at QA@CLEF 2004*, 2005.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. 1999.

[4] V. Jijkoun and M. de Rijke. A Pilot for Evaluating Exploratory Question Answering. In *Proceedings SIGIR 2006 workshop on Evaluating Exploratory Search Systems (EESS)*, 2006.

[5] L. Moreno-Monteagudo and A. Suarez. Una Propuesta de Infrastructura para el Procesamiento del Lenguaje Natural. In *Proceedings of SEPLN 2005*, 2005.

[6] G. Puscasu. A Framework for Temporal Resolution. In *Proceedings of the 4th Conference on Language Resources and Evaluation (LREC2004)*, 2004.

[7] P. Tapanainen and T. Jaervinen. A Non–Projective Dependency Parser. In *Proceedings of the 5th Conference of Applied Natural Language Processing, ACL*, 1997.

[8] Antonio Toral and Rafael Muñoz. A proposal to automatically build and maintain gazetteers for named entity recognition using wikipedia. In *Workshop on New Text, 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 2006.