FBK-irst at CLEF 2007

Milen Kouylekov, Matteo Negri, Bernardo Magnini and Bonaventura Coppola
Fondazione Bruno Kessler FBK-irst, Trento, Italy
{kouylekov,magnini,negri,coppolab}@itc.it

Abstract

This report presents the outcomes of the activity carried out at FBK-irst for the participation in the CLEF-2007 Main QA track. Both the major improvements over last year's version of the DIOGENE system, and the results achieved in the evaluation exercise are reported.

Keywords

Question answering, Wikipedia, Anaphoric expressions processing

1 Introduction

The main novelties in this year's setting of the Main QA Task at CLEF are represented by:

- Introduction of topic-related questions. Questions, possibly referring to each other through anaphoric expressions, are organized into clusters related to a specific topic.
- Extended answer search space. Besides the past years document collection, Wikipedia articles were added as a possible answer source.

Even though the overall system architecture is the same we adopted for our previous participations to CLEF evaluation exercise (see [2]), some adaptations were necessary to address the increased complexity of this year's edition of the task. These are shortly overviewed in Section 2, which presents our work on the new answer search space, and Section 3, which reports on our simple approach to topic-related questions. Section 4 and 5 conclude the report respectively reporting the results achieved by DIOGENE in the CLEF-2007 Main QA task, and presenting directions for future work.

2 Exploring Wikipedia

This year the dataset provided by the organizers included a dump of the Wikipedia articles. The resulting new dataset posed new problems that had to be addressed, including:

- Processing Wikipedia articles.
- Integrating the new document source in an appropriate position in the DIOGENE system dataflow.

2.1 Processing Wikipedia Articles

Wikipedia articles contain different types of texts: information about a certain topic, formulas, lists, tables etc. We considered as a *processable unit* any text paragraph inside an article, apart form the Wikipedia links. Thus, we didn't process any other information that is contained in the other parts of the Wikipedia articles. For each *processable unit* we cleaned the text, using regular expressions, to remove the following text formatting information:

- HTML tags.
- Wikipedia Links
- Wikipedia Comments

As a result, the clean *processable units* were considered as potential answer sources. The open source search engine Lucene [1] was used to index these *Wikipedia documents*, while the MG search engine [5] has been used to index the news document collection as in the last year's version of the DIOGENE QA system.

2.2 Integration in the System Dataflow

We decided to integrate the Wikipedia document index inside the document retrieval component of DIOGENE. The system uses a document retrieval technique based on query relaxation loops [3]. Such technique is designed to output a limited set of ranked documents (at least 30, at most 100). The Wikipedia document collection, however, is only considered as an auxiliary information source due to the noisy documents it contains. Often, in fact, our first implementation of the cleaning procedure does not return fully reliable processable units. This is due to the large amount of unremoved tags, special symbols, or other XML annotations. As a result, Wikipedia documents are considered as a less reliable information source and are accessed only if an insufficient number of articles (less than 30) is returned by the MG search engine accessing the news document collection.

3 Dealing with Topic-Related Questions

The other new problem that we had to address was handling a set of questions which share the same $focus$. To handle this problem the $focus$ of the first question has be recognized. For this purpose, we adopt the following simple heuristic, which defines the $focus$ of a question as *the first noun phrase or multi-word after the main verb of the question, if it is capitalized, or the second if the first one is in lower case.*

Examples of the focus identified for some CLEF-2007 questions are the following:

1. *Question* – In quale anno é uscito il film Flashdance?
   (*In what year Flashdance came on the screen?*)
   Focus – Flashdance

2. *Question* – Quali sono i Grandi Laghi africani?
   (*What are the Great African Lakes?*)
   Focus – Grandi Laghi africani

3. *Question* – Chi é l'autore del libro "Giorni giapponesi"?
   (*Who wrote the book "Giorni giapponesi"?*)
   Focus – libro "Giorni giapponesi"

Once the $focus$ of the input question Q_1 is identified, it is added as a keyword (or a conjunction of keywords) to the search queries of the following questions Q_2, ..., Q_n in the cluster, unless it is already present among their terms.

4 Results

Apart from these slight modifications to the system's architecture, our submission to this year's edition of the CLEF QA task (results are reported in Table 1) has been obtained with the same system's components described in our previous participation in CLEF [2], and reflects the "work-in-progress" situation of the DIOGENE QA system.

task	*Overall (%)*	*Def. (%)*	*List (%)*	*Factoid (%)*	*Temp. (%)*
Italian/Italian	11.50	2.36	0.00	15.17	12.50

Table 1: System performance in the QA tasks

A preliminary analysis of the results achieved focused on the impact of the adaptations of the system to this year's task.

As for wikipedia articles, potential answer candidates have been extracted from such additional resource only for 9 questions (for a total of 38 candidates). Out of them, the final answer returned by DIOGENE came from Wikipedia in 6 cases, but only in one case it was the correct one (*i.e.* Q-0134: "Quanto dista Dunleary da Dublino" - "*How far is it from Dunleary to Dublin*").

As for topic-related questions, our focus extraction heuristic has been applied for 67 questions. The focus has been correctly added to the search keywords of a question in 42 cases, leading to 5 questions correctly answered. In 1 case it is not clear what the focus actually is, making a decision about its correctness rather difficult. This is:

Q-0113 – Qual é la capitale di Rhode Island?
(*What is the capital of Rhode Island?*)
Q-0114 – Dove si trova?
(*Where is it located?*)

5 Conclusions

In this report we presented our adaptations of the FBK-irst DIOGENE QA system, made to participate in the CLEF-2007 Main QA track. Such improvements addressed the problems posed by the two novelties of this year's edition of the task, namely the introduction of Wikipedia articles to extend the document collection, and the introduction of topic-related questions. The results achieved by the system show that our basic procedures dealing with such problems need to be refined. In particular, as a first step, the cleaning procedure designed to extract reliable processable units from Wikipedia articles will be improved, allowing for a more effective exploitation of such resource. As for topic-related questions, future improvements will address the focus selection strategy, either with refined heuristics, or with supervised approaches as proposed in [4].

References

[1] Erik Hatcher and Otis Gospodnetic. *Lucene in Action (In Action series)*. Manning Publications, December 2004.

[2] Milen Kouylekov, Matteo Negri, Bernardo Magnini, and Bonaventura Coppola. Towards Entailment-based Question Answering: ITC-irst at CLEF2006. In *Cross Language Evaluation Forum (Clef-2006)*, Alicante, Spain, 2006.

[3] Bernardo Magnini, Matteo Negri, Roberto Prevete, and Hristo Tanev. Is It the Right Answer? Exploiting Web Redundancy for Answer Validation. In *Proceedings of the 40th Annual Meeting*

of the Association for Computational Linguistics, (ACL-2002), pages 1495–1500, Philadelphia (PA), 7-12 July 2002.

[4] Matteo Negri and Milen Kouylekov. ”Who Are We Talking About?” Tracking the Referent in a Question Answering Series. In *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2007)*, Lagos, Portugal, March 29-30 2007.

[5] Ian H. Witten, Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images.* Morgan Kaufmann Publishers, San Francisco, CA, 1999.