A Fast Forward Approach to Cross-lingual Question Answering for English and German

Robert Strötgen, Thomas Mandl, René Schneider

University of Hildesheim, Information Science
Marienburger Platz 22 - 31141 Hildesheim, Germany
D-31141 Hildesheim, Germany

mandl@uni-hildesheim.de

Abstract

This paper describes the development of a question answering system for mono-lingual and cross-lingual tasks for the languages English and German. We developed the question answering system from a document and retrieval focused perspective. The system consists of question and answering taxonomies, named entity recognition, term expansion modules, a multi-lingual search engine based on Lucene and a passage extraction and ranking component. The overall architecture and heuristics applied during development are described. We discuss the results at CLEF 2005 and show potential future work.

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

Question answering, Named entities

1 Introduction

The question answering (QA) system developed at the University of Hildesheim for the participation in this years' QA track at CLEF[1] is mainly based on experience from multi-lingual retrieval in previous years. Our system can do mono-lingual QA and cross-lingual retrieval, both for German and English as topic and document language. The architecture of this basic QA system is based on a retrieval engine developed for multi-lingual ad-hoc retrieval (Hackl et al. 2005). Further components necessary for a QA system (Harabagiu & Moldovan 2003) and some for system improvement were developed additionally.
As required components we implemented a question and answer taxonomy, a translation utility for automatically translating questions and a passage extraction and ranking passages from the documents. In addition, we integrated a tool for named entity recognition and term expansion. Many of the components were developed within a class for graduate students. All source code was developed with JAVA.

2 Query Processing

The query processing includes the assignment of a question and expected answer type, named entity recognition, translation and stopword removal.

[1] http://clef-qa.itc.it/

2.1 Question and Answer Taxonomies

A question taxonomy based on the questions of previous QA tracks (Magnini et al. 2005) was developed. It contains eleven question classes and several subclasses for the question types WHO, HOW, WHAT and WHERE and the corresponding answer classes.
An evaluation based on the CLEF QA topics form the years 2003 and 2004 showed that overall, for 73% of the questions, the answer category was assigned correctly. For further 14%, the categorization was partly correct and for another14% of the questions, a wrong category was assigned. The taxonomy was most reliable for the question types WHEN, WITH WHAT and FOR WHAT. Questions starting with WHAT were categorized worst.

2.2 Named Entity Recognition

Previously, we analyzed the impact of named entities on query performance in ad-hoc retrieval and found, that queries are often solved better when named entities are present (Mandl & Womser-Hacker 2005). As a consequence, we included named entity recognition from the beginning. The goal was, to identify named entities and to create a separate index for them. An analysis of three named entity recognition systems on the CLEF topics showed that the performance is satisfying and can be improved by training (Mandl et a.. 2005).
LingPipe[2] was used as a basic tool. Lingpipe applies a statistical machine learning approach to named entitiy recognition and categorization. For training LingPipe, we used one annotated corpus for each language:

- German: Frankfurter Rundschau with 36 Million word forms (Source: Linguistic Data Consortium, LDC[3])
- English: Reuters News (810.000 news texts)

An evaluation revealed a recognition rate of 60% for correct recognition and 42% for correct categorization into the following four classes: Person (PER), Organization (ORG), Place (LOC) und Miscellaneous (MISC).
Named entity recognition was applied to the queries and to the document corpus.

2.3 Query Translation

The key component for cross-lingual QA is a translation utility. As underlying systems, we used Babelfish, FreeTranslation and Linguatec[4]. To avoid a large influence of wrongly translated named entities, we replaced all named entities found in the query except for the category MISC with a dummy which was not translated by the translation tools. In addition, the named entities were sent to the translation tool without context subsequently. All translated sentences and terms were collected and only stopwords were removed.

2.4 Term Expansion

For retrieving German answers, the translated keywords were expanded using GermaNet[5]. However, to avoid the addition of too many senses, the expansion was only carried out, when GermaNet included only one meaning of the word under question. For English, the synonym function of WordNet[6] was used to expand all translated terms. The effect of term expansion has not been evaluated for our system yet.

3 Searching and Passage Retrieval

For stemming, indexing and retrieval we employed Lucene[7] as it has been used in (Hackl et al. 2005). The system searched with the keywords provided and first returned documents. These were split into passages of size of at least 200 including the remainder until the next punctuation mark.

---

[2] http://www.alias-i.com/lingpipe/
[3] http://www.ldc.upenn.edu/
[4] http://babelfish.altavista.com/, http://www.freetranslation.com/, http://www.linguatec.net/online/ptwebtext/
[5] http://www.sfs.nphil.uni-tuebingen.de/lsd/
[6] http://wordnet.princeton.edu/doc
[7] http://lucene.apache.org/

These passages were again indexed as documents by Lucene and ranked according to a scoring algorithm which rewards the frequency of occurrence of keywords in the passage (Light et al. 2001). The same set of keywords was used for retrieval and ranking. The top ranked passages are returned. A user interface which allows question input and which shows the top three passages has also been developed.
A few heuristics were implemented to improve performance. We focused on named entities especially.

- If named entity is the expected answer type and there are documents in the answer set which contain named entities of the appropriate type, then only these documents are forwarded to the passage extraction.
- If named entity is the expected answer type the most frequent named entities of the expected type within all passages are determined and the first passages containing these named entities are returned.
- If no answer with named entities is found, then the first 90 characters of the most highly ranked passage are returned.
- Trivial answers are not returned. Answers are considered trivial if they contain only one word, if they consist in the name of a known news agency of if the answer string is a subset of the question string.
- When the expected answer type is named entity, then all named entities in the first 20 passages are extracted and the most frequent named entity is returned.

The confidence weight returned by the system is the retrieval status value returned by Lucene for the returned passage. NIL is returned when no document is found by Lucene and in this case, a confidence value of 1.0 is assigned.

4 Experiments and Results

The quality of the results was only satisfying for definition questions. For this first participation and considering the focus on named entities, this seems acceptable. The results are shown in table 1.

Table 1. Results for QA system of the University of Hildesheim in 2005

Languages	Question Type	Accuracy
English -> German	Definition	18.00%
English -> German	Factoid	0.83%
English -> German	All	5.00%

The weak performance is probably due to several reasons. The time and effort dedicated to evaluation was mainly aimed at system stability and the integration of all tools. Parameter tuning based on previous CLEF experiments were not carried out so far. In addition, this year CLEF required a very short answer. Our system returns passages of at least the length 200 and no further processing is done to extract a short answer. This was probably an advantage for our system for definition questions, where the performance was good.

5 Outlook

The system for QA can be improved by further integrating the question analysis and the search process. So far, the knowledge gained from the question in not fully exploited. Furthermore, the system needs to be evaluated more thoroughly.

Acknowledgements

We would like to acknowledge the work of several students from the University of Hildesheim who implemented the components of the QA system as part of their course work[8].
We also want to thank Maarten de Rijke for his comments on an earlier version of our QA approach.

[8] http://www.uni-hildesheim.de/~rschneid/psws04odqa.html

References

Hackl, René; Mandl, Thomas; Womser-Hacker, Christa (2005): Mono- and Cross-lingual Retrieval Experiments at the University of Hildesheim. In: Peters, Carol; Clough, Paul; Gonzalo, Julio; Kluck, Michael; Jones, Gareth; Magnini, Bernard (eds): Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign. Berlin et al.: Springer [LNCS 3491] pp. 165 – 169.

Harabagiu, Sanda & Moldovan, Dan (2003): Question Answering. In: *The Oxford Handbook of Computational Linguistics*. Oxford; New York: Oxford University Press, 2003.

Light, Marc; Mann, Gideon S.; Riloff, Ellen; Breck, Eric (2001): Analyses for elucidating current question answering technology. In: Journal of Natural Language Engineering, Special Issue on Question Answering Fall-Winter 2001.

Magnini, Bernardo; Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Simov and Richard Sutcliffe (2005): Multiple Language Question Answering (QA@CLEF). Overview of the CLEF 2004 Multilingual Question Answering Track. In: Working Notes 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004. Bath, England, http://clef.isti.cnr.it/2004/working_notes/WorkingNotes2004/35.pdf

Mandl, Thomas; Schneider, René; Schnetzler, Pia; Womser-Hacker, Christa (2005): Evaluierung von Systemen für die Eigennamenerkennung im cross-lingualen Information Retrieval. In: Gesellschaft für linguistische Datenverarbeitung. Beiträge der GLDV-Frühjahrstagung. Bonn, 30.3. - 01.04. Frankfurt a. M. et al. Peter-Lang.

Mandl, Thomas; Womser-Hacker, Christa (2005): The Effect of Named Entities on Effectiveness in Cross-Language Information Retrieval Evaluation. In: Proceedings ACM SAC Symposium on Applied Computing (SAC). Information Access and Retrieval (IAR) Track. Santa Fe, New Mexico, USA. March 13.-17. 2005. S. 1059-1064.