Learning Methods in Multi-grained Query Answering*

Philipp Sorg

Institute AIFB, University of Karlsruhe
D-76128 Karlsruhe, Germany
sorg@aifb.uni-karlsruhe.de

Abstract. This PhD proposal is about the development of new methods for information access. Two new approaches are proposed: Multi-Grained Query Answering that bridges the gap between Information Retrieval and Question Answering and Learning-Enhanced Query Answering that enables the improvement of retrieval performance based on the experience of previous queries and answers.

1 Introduction

Finding relevant information in the WWW, in knowledge bases of companies, in document repositories or even on personal computers is getting more and more important, as the amount of knowledge contained in these resources continuously increases. In addition, users in a private or professional environment rely heavily on the information. In a professional environment, e.g. as described by Abecker et al [1], building and using Organisational Memories is essential for all companies working in the information sector and reducing the effort of finding information is an important cost factor.

In my PhD research I plan to develop new methods for searching in such heterogeneous knowledge bases. In this PhD proposal I will describe current search methods, identify missing features and present new ideas to improve current search systems.

1.1 Motivation

Current Information Retrieval (IR) and Question Answering (QA) systems traditionally only return answers of a specific granularity. While there are some exceptions, e.g. the search engine Google[1] that implements some heuristics to detect and answer simple factoid questions, IR systems typically return whole documents as answers. On the other end of the spectrum, QA systems try to find an exact answer to the question. They achieve reasonable retrieval results

* This work was funded by the Multipla project sponsored by the German Research Foundation (DFG) under grant number 38457858. Many thanks to my PhD supervisor Dr. Philipp Cimiano for his helpful comments.

[1] http://www.google.com

A. Sheth et al. (Eds.): ISWC 2008, LNCS 5318, pp. 926–931, 2008.

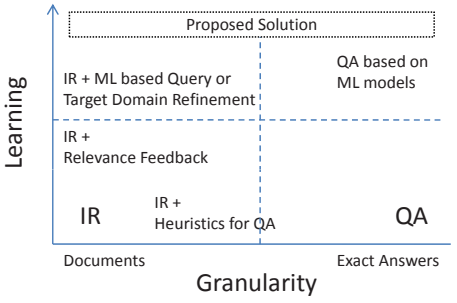

Fig. 1. Current Information Retrieval and Question Answering methods, classified by answer granularity and involved learning

on factoid questions, but are in general not able to answer complex questions, e.g. questions based on a large context or questions for which the answer is not explicitly stated in the text but must be inferred. The gap between IR and QA systems could be filled by systems that handle answers of different granularity ranging from whole documents to exact answers in a flexible way.

Another missing feature of most current IR and QA systems is the missing ability to learn from experience. Intuitively, systems should be able to use information extracted from previous pairs of queries and answers and use this information to improve retrieval results. As far as we know there is no prominent search system that supports this kind of learning [2].

1.2 Current State of My PhD Research

I started my PhD research in October 2007.

First I started to examine different methods of Natural Language Processing (NLP), e.g. relatedness measures on terms and text. A special focus was on defining semantic relatedness measures based on Wikipedia, e.g. by using Explicit Semantic Analysis (ESA) [3] that represents text in a Wikipedia article space. As part of this research I developed a new method to learn new cross-language links in Wikipedia [4], which I used in cross-lingual ESA to define a relatedness measure across languages.

At the current stage I am developing a general framework for accessing and processing unstructured (e.g. plain-text documents) and structured (e.g. ontologies) information sources. Based on this framework I plan to implement the new query answering methods presented in this PhD proposal.

2 Definition of the PhD Topic

In the following section I will define the problem I intend to address during my PhD research and describe state of the art methods that address this problem.

2.1 Definition of the Problem

The problem I want to investigate is to develop new search methods supporting the following features:

- Detect the right granularity of answers and return answers of different granularity. *[Multi-Grained Question Answering]*
- Learn from previous queries and answers to improve retrieval results. *[Learning-Enhanced Question Answering]*

2.2 State of the Art

The problem of Multi-Grained Query Answering is to some extent addressed by commercial Internet search engines like Google or Yahoo where the answer space is mainly defined at the document level but small possibly relevant text snippets are presented as well. Another approach is to use supervised learning methods to learn the ranking function that is used to retrieve document elements [5].

The problem of answering factoid queries based on background knowledge is e.g. solved by matching the query to certain patterns (as implemented in Google) or by finding relevant text by using IR methods on word level and pinpointing the right answer using linguistic analysis on a syntactic/semantic level [6].

A method to improve the retrieval system using previous queries and answers is to use relevance feedback from users [7]. Another approach is to use Machine Learning models trained on query-answer pairs to translate query terms to answer terms for target domain refinement [8].

An example for a QA system based in Machine Learning can be found in [9], where queries and answers are represented as graphs and graph rules mapping queries to answers are learned, which are used for the QA system. Another approach is to learn patterns from question/answer pairs that can be used for QA. A bootstrapping pattern mining approach is e.g. described in [10], where starting from a few hand-crafted examples new patterns are inferred from the Internet using a web search engine.

3 Approach to the Problem

In this section I will first describe how I intend do analyse existing retrieval methods, ranging from Information Retrieval to Question Answering. Then I will present initial ideas for Multi-Grained and Learning-Enhanced Query Answering.

3.1 Analysis of Existing Retrieval Methods

The analysis of existing retrieval methods will focus on IR and QA methods. The expected outcome will be an overview of current retrieval methods and the identification of strengths and weaknesses of those methods.

IR Methods. The analysis of existing IR methods will be mostly concerned with vector space representations of text. The most simple representation is the standard Bag-of-Words model, but there are many systems that extend this model with different weights, by using similarity measures to deal with synonyms or by including background knowledge like annotations of Named Entities. This analysis provides the foundation of purely statistical approaches of Query Answering.

Another important aspect is the use of relevance feedback in IR systems. This is often done by using Machine Learning techniques and will therefore be substantial for the Machine Learning part of my research.

QA Methods. QA systems normally use a more structured representation of text, often based on deep linguistic analysis. A big variety of background knowledge is used in current systems, ranging from patterns matching factoid answers to complex ontologies. This analysis will help to find appropriate representations of text that can be used to develop new retrieval methods.

Many QA systems use IR methods to identify relevant parts of documents. The analysis of these methods will be important as Multi-Grained QA will be based on these existing methods.

3.2 Description of Envisioned Methods

Multi-grained Query Answering. The core of Multi-Grained QA is to develop methods that are able to identify the right granularity of the answers given a query and based on the information sources. One idea is to introduce a measure of Answer Density. The trade-off between completeness of the answer and its length should be modeled in this measure.

The following example shows the advantage of such an Answer Density measure to existing QA systems. For the question

> Why did David Koresh ask the FBI for a word processor?

it is not possible to determine an expected answer type. Users asking this question would expect a short paragraph containing an explanation, like this text snippet of the Wikipedia article "David Koresh":

> ... Communication over the next 51 days included telephone exchanges with various FBI negotiators.
> As the standoff continued, Koresh, who was seriously injured by a gunshot wound, along with his closest male leaders negotiated delays, possibly so he could write religious documents he said he needed to complete before he surrendered. ...

Ideally the Answer Density measure would assign a high value to this snippet. This could e.g. be done by using the semantic relatedness of "word processor" and "write". As the presented snippet is part of the article "David Koresh", contains the term "FBI" and is related to "word processor", the value of the Answer Density is high and could be identified as possible answer.

Learning-Enhanced Query Answering. The problem of learning from previous queries and answers is a problem of unsupervised ML. Supervised learning methods based on user feedback yield good results in improving retrieval performance, but have the problem that feedback is not available in general. As these methods also are widely discussed in the IR research community I intend to focus on unsupervised ML techniques.

One idea is to use clustering techniques to cluster queries and answers. Based on this clustering, query-query, query-answer and answer-answer relations can be extracted. We plan to use syntactic and semantic features of the query for the clustering.

The syntactic features can be used to identify the expected type of answer. E.g. if a question starts with "Who ..." the expected answer will probably be factoid, whereas a more detailed answer is needed for questions starting with "Why ...". Clustering based on Syntactic Tree Kernels [11] is a possible method to use the syntactic features of the queries.

Semantic features express the topic and content of the query. We plan to use Wikipedia as background information source by mapping queries to a space of Wikipedia articles using extracted Named Entities that correspond to Wikipedia articles. This mapping can be used to identify queries with similar topics. Extracted key terms from these queries can then be used for query or answer refinement.

Another application of this mapping to the space of Wikipedia articles is the usage of the categories of these articles. Combined with syntactical information a more abstract representation of queries and answers can be constructed. It is then e.g. possible to use the categories of these articles together with the syntax of the query to find an abstract representation that can be used to cluster similar queries. E.g. the question *"Who wrote Faust?"* with the factoid answer *"Goethe."* could be represented as *"Who wrote* $\{Book\}$*?" "*$\{Person\}$*."*. For new queries assigned to the same cluster the system can then infer that the answer should contain a person.

4 Evaluation

There are different approaches for the planned evaluation of the developed methods. One is automatic evaluation based on existing datasets. As there are no existing datasets for Multi-Grained Query Answering, this evaluation can only be applied to the extrema of Multi-Grained Query Answering by using datasets for the evaluation of IR or QA systems, e.g. datasets provided by TREC[2]. As it will probably not be possible to use this evaluation to compare the new methods with existing IR or QA methods due to the differences in the answer granularity, this evaluation can be mainly used to analyse the benefit of Learning-Enhanced QA. After learning the results should improve on the used datasets. One evaluation step could be the comparison of results of the same query before and after the training phase.

[2] `http://trec.nist.gov/`

To compare the developed system with other IR or QA systems I plan to perform a manual evaluation based on a real world scenario, e.g. a user evaluation involving several people using the system as a desktop search engine.

5 Conclusion

I have presented my PhD research proposal in the field of IR/QA to enhance information access. The main goal is to overcome the rigidity of current systems which either only return full documents or try to pinpoint exact answers. Further, I aim at developing paradigms by which systems can learn from past experience which represents a crucial open problem in the field of information access.

References

1. Abecker, A., Bernardi, A., Hinkelmann, K., Kuhn, O., Sintek, M.: Toward a technology for organizational memories. IEEE Intelligent Systems 13(3), 40–48 (1998)
2. Strzalkowski, T., Harabagiu, S.: Advances in Open Domain Question Answering (Text, Speech and Language Technology). Springer New York, Inc., Secaucus (2006)
3. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (2007)
4. Sorg, P., Cimiano, P.: Enriching the crosslingual link structure of wikipedia - a classification-based approach. In: Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence, Chicago, IL, USA (2008)
5. Vittaut, J.N., Gallinari, P.: Machine learning ranking for structured information retrieval. In: Proceedings of the European Conference on Information Retrieval, pp. 338–349 (2006)
6. Hovy, E.H., Gerber, L., Hermjakob, U., Junk, M., Lin, C.Y.: Question answering in webclopedia. In: Proceedings of the Text Retrieval Conference (2000)
7. Harman, D.: Relevance feedback revisited. In: Belkin, N.J., Ingwersen, P., Pejtersen, A.M. (eds.) Proceedings of the Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, pp. 1–10. ACM, New York (1992)
8. Riezler, S., Vasserman, A., Tsochantaridis, I., Mittal, V., Liu, Y.: Statistical machine translation for query expansion in answer retrieval. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic (2007)
9. Molla, D., van Zaanen, M.: Learning of graph rules for question answering. In: Proceedings of the Australasian Language Technology Workshop, Sydney, Australia (2005)
10. Ravichandran, D., Hovy, E.: Learning surface text patterns for a question answering system. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, pp. 41–47. Association for Computational Linguistics (2001)
11. Moschitti, A.: Efficient convolution kernels for dependency and constituent syntactic trees. In: Proceedings of the European Conference on Machine Learning, Berlin, Germany, pp. 318–329. Springer, Heidelberg (2006)