Web-Based Multiple Choice Question Answering for English and Arabic Questions

Rawia Awadallah and Andreas Rauber

Department of Software Technology and Interactive Systems,
Vienna University of Technology, Vienna, Austria
rradi@mail.iugaza.edu, rauber@ifs.tuwien.ac.at

Abstract. Answering multiple-choice questions, where a set of possible answers is provided together with the question, constitutes a simplified but nevertheless challenging area in question answering research. This paper introduces and evaluates two novel techniques for answer selection. It furthermore analyses in how far performance figures obtained using the English language Web as data source can be transferred to less dominant languages on the Web, such as Arabic. Result evaluation is based on questions from both the English and the Arabic versions of the TV show "Who wants to be a Millionaire?" as well as on the TREC-2002 QA data.

1 Introduction

A large body of research exists on Question Answering (QA) where user queries are received in a natural language and precise answers are returned, decomposing the problem into three steps: (1) retrieving documents that may contain answers, (2) extracting answer candidates, and (3) selecting the most probably correct answer. Early TREC QA systems were looking for an answer that was known to be included in a given local corpus. Now, many QA systems use the Web as a corpus, either by extracting answers or by learning lexical patterns from the Web which are then used to improve the system itself. Studies suggest that the resulting data redundancy provides more reliable answer extraction [1]. Different approaches to improve system performance exist, such as using probabilistic algorithms to learn the best question paraphrase [2] or training a QA system to find possible sentence-length answers [3]. When several potential answers are retrieved, answer validation techniques rank them, selecting the most probable answer. This basically resembles multiple-choice QA. Approaches to answer validation range from purely statistical methods [7] based on Web search to the use of semantic techniques [4].

In this paper we present and evaluate two new answer selection techniques within a multiple-choice QA settings, comparing them to excisting answer validation techniques. These are evaluated on both English and Arabic language questions to evaluate the impact of the different sizes of the Web in the respective languages. Questions stem from both the TREC-2002 QA task questions as well as the English and the Arabic versions of the TV show "Who wants to be a Millionaire?", a quiz-show that originated in the UK and has been exported around the world, where candidates have to answer 4-choice trivia general-interest questions.

M. Lalmas et al. (Eds.): ECIR 2006, LNCS 3936, pp. 515–518, 2006.

The remainder of the paper is organized as follows: Section 2 describes our multiple-choice QA module. Experiments are detailed in Section 3, with conclusions being presented in Section 4.

2 The MCQAS Module

The core procedure of our Multiple Choice Question Answering System (MCQAS) is roughly as follows: A set of representative keywords both from the question and from each individual answer is extracted using simple linguistic techniques. Tokenization is performed to extract individual terms followed by (attached and detached) stop word and punctuation removal. The stem of each of the remaining words is obtained. For Arabic, a normalization process is further applied on the remaining words as described in [8]. The set of these remaining words along with their stems form the keywords set which is transformed into a set of individual queries combining the question keywords and the answer keywords of each individual answer. This is then submitted to, in our case, the Google search engine. A core task now is to assess the relevance of the candidate answers. Using search engines and the Web as a basis for answer selection, several different techniques utilizing different amounts of information can be applied. Those range from simple hit counts, via using the text snippets returned for each document providing context information on the query words found, to full-fledged analysis of the documents retrieved by the search engine. As the latter results in a rather high overhead in terms of document downloads, our work focuses on utilizing the result snippets for answer selection. In MCQAS, six answer selection techniques are used – four were previously used in answer validation task and two new ones. These are either based on the number of documents retrieved from the Web (Hits, CCP, KA), or on the analysis of snippets returned by the search engine (CW, AQC, AQA):

1. **Hits:** simple hit counts returned by a search engine [5].
2. **Corrected Conditional Probability (CCP):** based on the conditional probability of answer keyword based hits, given query keywords [7].
3. **Key Words Association (KA):** based on forward and backward associations of the query using hand crafted rules, calculating probabilities for hits using the set of question and answer keywords.
4. **Co-occurrence Weight (CW):** based on the distance (number of non-stopwords) between question and answer keywords in result snippets [7].
5. **Answer and Question words Count (AQC):** based on the number of question and/or answer keywords ocurring in result snippets.
6. **Answer and Question words Association (AQA):** based on the co-occurrence of both question and answer keywords within the same result snippet's context.

In a nutshell, the two new techniques are calculated as follows: The snippets of the first 10 (or all, if less than 10) search results for each query are weighted, and their average should be the answer score. For AQC, a snippet weight is the number of query words it contains. For AQA, a snippet weight is the sum of its sub-snippets weights where the sub-snippet (context) is defined by the text between the ellipsis symbols **"..."**, and in which at least one question keyword and at least one answer keyword co-occur. A sub-snippet weight is the percentage of the different question keywords added to the percentage of the different answer keywords.

3 Experiments

In order to check the validity of the different answer validation techniques experiments have been carried out using questions from the English[1] and the Arabic version of the TV Show "Who Wants to Be a Millionaire?", as well as the TREC-2002 QA track questions. To transform the latter into a multiple choice QA setting four answers returned during the TREC sessions were selected manually for each question, making sure that exactly one correct answer is among the four.

Table 1. QA accuracy of different techniques for different questions categories

Category	**Hits**	**CCP**	**KA**	**CW**	**AQC**	**AQA**
Arabic	38.0%	43.0%	45.0%	50.0%	44.0%	**55.0%**
English	43.0%	45.0%	48.0%	59.0%	**63.0%**	60.0%
TREC	35.0%	40.0%	42.0%	59.0%	**62.0%**	56.0%

A random subset of 100 questions was used to run the experiments in each case. An overview of the results is provided in Table 1. The snippet-based techniques outperformed the hits-based ones. For Arabic, AQA outperforms the other techniques, while for English, AQC is dominant. An analysis of the Arabic queries search results has revealed, that the returned number of snippets for most queries was less than 10 and most of these snippets were irrelevant and only few relevant precise phrases were found to exist on the Web. This is because there are many Arabic words with the same spelling but with different meanings. So the use of more restrictive schemas (CW and AQA) is essential. More over, using general search engines such as Google for Arabic queries does not satisfy the redundancy issue required by the hits-based techniques since Arabic specific features to query correction such as word morphology or word root is not implemented, which emphasizes the need for more linguistic efforts. On the other hand, for English queries the redundancy is higher and more restrictive schemas may ignore the cases where the question and the right answer keywords appear frequently but in different contexts (sub-snippets).

A more detailed analysis reveals that the various techniques tend to answer different questions correctly. This opens room for ensemble methods. However,

[1] Thanks to Shyong K. Lam for providing us with their test data from [5].

more detailed analysis of question types and answer characteristics will be required to reveal an optimized strategy.

4 Conclusions

In this paper we proposed two new techniques for answer selection based on analyzing the text snippets returned by a search engine when confronted with modified question–answer pairs as queries. Evaluations have been performed both on English and Arabic questions from the TV show “Who wants to be a Millionaire?” as well as TREC-2002 data. Experiments reveal an average performance of 55-62%, with the AQA strategy performing better on the Arabic language questions, while AQC is superior for English language tasks. This may be attributed to the morphological complexity of the Arabic language, resulting in only precise phrases returned if they exist on the Web, rather than having split segments returned as well. Analysis reveals that further improvements can be obtained by both more complex linguistic pre-processing, specifically for the Arabic language, and by using ensemble methods for answer selection.

References

1. Clarke, C. L. A. and Cormack, G. V. and Lynam, T. R., *Exploiting redundancy in question answering*, Proc. of th 24th Annual Int’l ACM SIGIR Conf. on Research and Development in Information Retrieval, 2001.
2. Radev, H.R and Qi, H. and Zheng, Z. and Blair-Goldensohn, Zhang, Z. and Fan, W.and Prager, J., *Web for Answers to Natural Language Questions*, Proc. of the 10th Int’l Conf. on Information and Knowledge Management, 2001.
3. Mann, S., *A Statistical Method for Short Answer Extraction*, Proc. of the 39th Annual Meeting of the Association for Computational Linguistics, 2001.
4. Harabagiu, S. and Maiorano, S., *Finding Answers in Large Collections of Texts: Paragraph Indexing + Abductive Inference*, Proc. of the AAAI Fall Symposium on Question Answering Systems, 1999.
5. Shyong, K. Lam and David, M. Pennock and Dan, Cosley and Steve, Lawrence, *1 Billion Pages = 1 Million Dollars? Mining the Web to Play “Who Wants to be a Millionaire?”*, Proc. of the 19th Conf. on Uncertainty in Artificial Intelligence, 2003.
6. Masatsugu, T. and Takehito, U. and Satoshi, S., *Answer Validation by Keyword Association*, Proc. of the 3rd Workshop on Robust Methods in Analysis of Natural Language Data, 2004.
7. Magnini, B. and Negri, M. and Prevete, R. and Tanev, H., *Mining the Web to Validate Answers to Natural Language Questions*, Proc. of the 3rd Int’l Conf. on Data Mining, 2002.
8. Ballesteros, L. and Connell, E.M., *Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis*, Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2002, 275-282.