Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 5, May 2013, pg.56 – 59

RESEARCH ARTICLE

Judging Relevance of Candidate Answers in Question Answering System

Rawia Awadallah[1]

[1]Computer Science Department- Islamic University of Gaza, Palestine

[1] *rradi@iugaza.edu.ps*

Abstract— Building a prototype for answering multiple choice general-interest trivia questions requires an approach for answer selection. In this paper, an approach for answer selection is proposed. The main objectives in this context are: to show that the World Wide Web can be used as data source for answering general-interest questions in a languages such as Arabic and to compare among the performance of different answer selection strategies and their performance with different languages.

The different experiments that have been carried out, have shown that the World Wide Web's redundancy enables the utilization of simple tricks to overcome many troublesome issues in natural language processing, such that for the English language, simple statistical techniques suffice to validate a set of candidate answers while more pre-processing linguistic efforts are still needed for the Arabic language.

It has been show that 61% and 56%. Of ''Who wants to be a millionaire'' English version questions and Arabic version questions can be solved by the proposed method respectively.

Key Terms: - Question Answering, Information Retrieval

I. 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 two steps: retrieving documents that may contain answers, and extracting precise answers from these documents. 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 answer [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 the use of semantic techniques [4] to purely statistical methods [7] based on Web search.

We thus evaluate two new answer selection techniques within a multiple choice QA settings. Furthermore, we test the scalability of the existing answer validation techniques to these settings. 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 version of the TV show ”Who wants to be a Millionaire?”; a show originated in the UK and has been exported around the world and depends on the ability to answer 4-choice trivia general-interest questions.

The remainder of the paper is organized as follows. Section II describes our multiple-choice QA module. Experiments are detailed in Section III, with conclusion being presented in Section IV.

II. The MCQAS Module

One basic task of the Multiple Choice Question Answering System (MCQAS) is to assess the relevance of a candidate answer with respect to a given question. For instance, given the question **"What is the capital of Egypt? "**, the relevance of the answer **"Cairo"** is equivalent to estimating the correctness of the query: **"The capital of Egypt is Cairo"**. If this query is sent to a search engine such as Google, a set of passages is returned representing the documents where the query words occurred. These passages snippets contain a significant amount of knowledge about the relations between the question and the answer.

A common feature in these snippets is the occurrence of a certain subset of words (i.e **"capital"**, **"Egypt"** and **"Cairo"**). Some of those snippets are lexically similar to the question and the answer and some are not similar. If the query covers a large portion of snippets, then the answer in this case can be considered to be relevant and vice versa. A useful feature of such query is that when we search for it on the Web it usually produces many hits, thus making statistical approaches and content approaches applicable.

Starting from the above considerations and given a question-answer pair, MCQAS applies an answer selection procedure based on the following general steps:

1. *The set of representative keywords from the question* ***QK*** *is computed. This step is carried out using linguistic techniques (Stop words removal, stemming [8], normalization, etc.).*
2. *From the extracted keywords the query for the pair the answer word* ***"a"*** *and the question's keywords* ***QK*** *is computed.*
3. *The query is submitted to the Web and the relevance of each candidate answer is estimated.*

A. Answer and Question Words Association Measure

MCQAS basic strategy to estimate answer relevance is based on the co-occurrence of question keywords and an answer's words in search results snippets. The first 10 (or all, if there were less than 10) snippets Google returned for each query are analysed and weighted as follows:

Having co-occurrence of a set words of an answer $A = \{a_1, a_2, \ldots\}$ and a set of keywords of a question $QK = \{q_1, q_2, \ldots\}$, the **sub-snippet weight** $w(A, QK)$ for each result sub-snippet; defined by the text between ellipsis symbol \textbf{"..."} and in which at least one question keyword and at least one answer word co-occur; is calculated by means of the following formula:

$$w_i(A, QK) = \frac{|QK_i|}{|QK|} + \frac{|A_i|}{|A|}$$

Where $|QK_i|$ is the number of different question keywords that occur in sub-snippet i, while $|A_i|$ is the number of different answer words that occur in that sub-snippet.

The **snippet weight** $W(A, QK)$ is computed using the following formula:

$$W(A, QK) = \sum_i w_i(A, QK)$$

The **answer score** $S(A)$ formula is used to score each candidate answer as follows:

$$S(A) = \frac{\sum_k W_k(A, QK)}{k}$$

Where $k \leq 10$ represents the set of the first ten snippets (or less) where any answer word in A appears in.

The response to the question is normally the answer that produces the highest score. However, a number of questions were identified by the presence of the word "not" in the question. In such cases, the answer yielding the fewest results is chosen.

III. Experiments and Results

In **MCQAS**, six different answer selection strategies have been implemented for the purpose of comparison:

1. *Hits Strategy: A base line strategy based on number of hits returned by Google search engine [5].*
2. *Corrected Conditional Probability (CCP) Strategy: A strategy based on the asymmetry of the question-answer relation [7].*
3. *Key Words Association (KA) Strategy: A strategy based on the forward or backward association of the conjunction query [6].*
4. *Co-occurrence Weight (CW): A strategy based on the number of question keywords and their distance from a candidate answer in search results snippets [7].*

5. *Answer and Question words Count Strategy (AQC): A strategy based on the occurrence of question and answer words in search results snippets.*
6. *Answer and Question words Association (AQA): A strategy based on the co-occurrence of question and answer words in search results snippets (* (A) *score as described in previous section).*

Some of these strategies are based on the statistical approach- Hits, CCP, and KA- and the other are based on the content-based approach- AQC, AQA, and CW.

TABLE I

QA accuracy of different techniques for different questions categories

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

In order to check the validity of the different existing answer validation techniques besides the two new techniques for our answer selection task, experiments have been carried out using questions from the English and the Arabic version of the TV Show ”Who Wants to Be a Millionaire?”, as well as the TREC-2002 QA track questions. For the latter, four answers returned during the TREC sessions were selected manually for each question; making sure that exactly one correct answer is among the four; to transform it into a multiple choice QA setting. 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 returned if they 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. Moreover, 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 satisfied 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 further analysis of the results has shown that among the different strategies, at least one has the correct answer for about 85%, 87% and 72% in the 100-question sample of the English Millionaire questions, in TREC 100-question sample and in the 100-question sample of the Arabic Millionaire questions respectively.

This reveals that the combination of two or more strategies may increase the correctness percentage. To exploit the reliability of this issue and to construct the corner stone for a future work in this topic, the experiment's results of 100-question sample of TREC set have been further analysed to find the relationship between any two strategies. See Figure 1. In all the relationships reported in this figure, it can be noticed that there are three main factors: the percentage of the questions that can be answered collectively by the two strategies, the percentage of the questions that can be answered correctly by both strategies, and the percentage of questions answered correctly by one strategy while wrongly answered by the other strategy. Selecting among these three factors is difficult.

But in general, it can be said that it is important when selecting any two strategies to combine, to put in consideration, that the two strategies can collectively answer correctly as large as possible number of questions. At the same time, these two strategies should both agree to answer correctly as large as possible number of questions. Moreover, it is important to notice the number of questions correctly answered by one strategy while wrongly answered by the other strategy. The larger these numbers, the more the need to combine these two strategies. This emphasizes the need to study the degree of confidence of each strategy which determines how much we believe in the judgement of one strategy compared to the judgement of another strategy, and to study the types of questions that a strategy usually correctly answers while the other strategy usually wrongly answers them. This initial analysis reveals that the various techniques tend to answer different questions correctly. This opens room for ensemble methods. However, more detailed analysis of question types and answer characteristics will be required to reveal an optimized strategy.

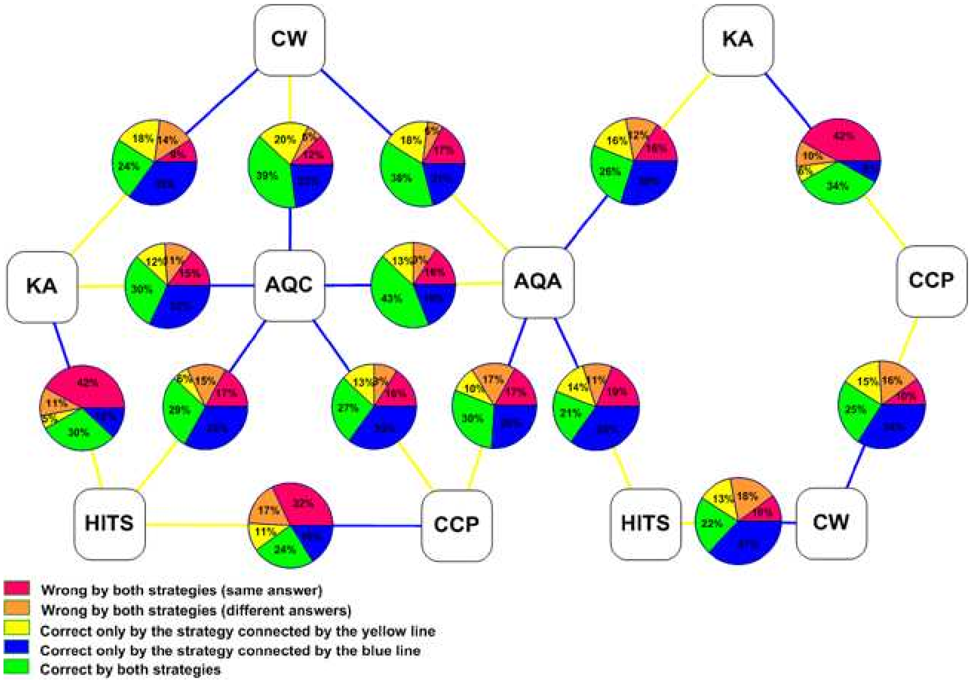

Figure 1 Question-answering accuracy compared among the different answer-selection strategies

IV. 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 ACQ 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.