Question Answering System Using Query Expansion and Heuristic Features

Yolanda Sjafrizal, Indra Budi, MirnaAdriani andPhilip Arthur

Information Retrieval Laboratory Lab, Faculty of Computer Science, Universitas Indonesia
Depok, Indonesia

yolanda.sjafrizal@ui.ac.id,indra@cs.ui.ac.id,
mirna@cs.ui.ac.id, philip.arthur@ui.ac.id

Abstract: This paper describes the design of question answering system that participates in the maintask of QA4MRE at CLEF 2013. This system will initially perform preprocessing stage of the document and the documents related questions. Then, it identifies the type of questions in order to be able to search the answers with the most appropriate approach. In order to finding the answers, the system uses eight heuristics features and two query expansion techniques, namely Pseudo Relevance feedback and WordNet as the lexical database. There are nine types of runs were submitted at CLEF 2013, where the difference lies in the combination of query expansion techniques and features used. Evaluation will be based on the c@1 measure. Best results are obtained when it uses both query expansion techniques and all features with certain weights.

Keywords: Question answering system, query expansion, pseudo relevance feedback, WordNet.

1 Introduction

CLEF (Conference and Labs of the Evaluation Forum) is an organization that plays an important role in improving research and development in the field of information retrieval, by providing the infrastructure for a number of activities as well as a series of lab evaluations on information systems and a number of conferences. One lab evaluation that conducted by CLEF is QA4MRE (Question Answering for Machine Reading Evaluation), with the main objective is to develop a best methodology for machine reading system, which will be evaluated through a process of questioning and testing the level of understanding of a reading [1]. Based on these issues, we conducted a number of experiments to develop a question answering system based on a dataset of QA4MRE in 2011 and 2012. The main task of this system is to conduct a thorough understanding of the single document provided as reading material, and then answer a number of questions related to the document. Questions are provided in the form of multiple choices, which each have 5 answer choices.

This paper proposed two query expansion techniques, Pseudo Relevance Feedback [2], and using WordNet as the lexical database. Beside of that, it is also uses eight heuristics features as that will give a score for each answer choice. To support the process of finding answers, the system also uses REVERB[1]to obtain the relation of the phrases of documents and snippets from Google snippets as additional knowledge.

In section 2 below, we describe the condition of dataset of QA4MRE at CLEF 2013, that used by this question answering system. Then, further explanation of the stages of the process will be described in section 3. The results of the system on a dataset of 2013 are described in section 4.

2 Dataset QA4MRE at CLEF 2013

The dataset provided by QA4MRE at CLEF 2013, consisting of four main topics, namely AIDS, climate change, Music & society and Alzheimer. A total of four pieces of reading test are provided at each such topic, which a reading test also consists of a single document with 10 questions and 5 answer choices for each question. So in general, for the dataset of 2013 there will be:

- 16 test documents (4 documents for each topic)
- 240/320 questions (15/20 questions for each document)
- 1200/1600 choice answers (5 answer choices for each question)

Dataset of 2013 is provided in a number of languages, like Arabic, Bulgarian, English, Romanian and Spanish. For this experiment, its scope was limited to the dataset that use English only.

3 Steps of The System

3.1 Pre-Processing

We conducted pre-processing on the dataset (document and the questions). The process are consists of 10 standard stages in the natural languages, such as convert the format of all numbers into a standard format, tokenization, named-entity recognition**,** filtering all punctuation characters that are considered useless, lowercasing**,** anaphora resolution**,** number recognition**,** lemmatization**,** removal of stopword, and processing negation. In negation processing, every word that has a meaning of negation, such as "not present" (containing the word "not"), changes to form of "not-present". We do this process so that the system can distinguish the word with its negation.

[1]https://github.com/knowitall/reverb-core/

3.2 Question Identification

At in this step, we identify the types of questions which are divided into factoid and non-factoid. The identification process based on the words that contains in the question, such as “what”, “why”, “where”, “who”, “when”, “how”, and “do”, as well as the identifier at the right of the question words (if any), such as countries, time, year, many, and etc. In addition, there are also questions that do not contain the question word, so it is more shaped like a command line. Sample questions as follows:

- “In what year she went to England? There is a question word "what" and identifier "year", so the question is identified as type of DATE.
- “Who is the president?”There is a question word "who", so the question is identified as a type of PERSON.

Having this type of question has been identified; next, the question will be given tags to indicate what kind of answer is expected from the question. We uses eleven tags, which are: "DATE", "LOCATION,"PERSON","OPINION","NUMBER","YES_NO","MONEY","CAUSAL","METHOD","PURPOSE", and "WHICH_IS_TRUE ".

3.3 Passage Retrieval

Passaging process will form a small part or a unit of text that called passage, wherein each passage consisting of *m* pieces of sentences. As example, in the text there are n sentences,S1, S2, S3, ...,Sn. Of the entire sentence will be formed a passage p, each consisting of 6 sentences, where

$$p = \{S_1 + S_2 + S_3 + S_4 + S_5 + S_6\} \quad (1)$$

So from the passaging process, will obtain a collection of passage P, with the formula:

$$\mathbf{P} = \{p_1 + p_2 + p_3 + \ldots + p_k\} \quad (2)$$

where k is number of passage that is formed from a text, m is number of sentences in a passage and n is number of sentences in the text. After the passaging process has been successfully carried out, we do the indexing process on the passages. We use Lucene[2] [3] to create the inverted-index.

After that, we do the retrieval process, getting the passages that contain the answers. In this system, we assumed that the top 13 passage that has the greatest similarity value (most relevant) to the query will be taken. This figure is obtained after seeing the accuracy of the system, when conducted experiments using the numbers 1 to 15. After that, the system looks for the appropriate answer based on the relevant passages retrieved. When it finds no answer to the question, then the system will perform query expansion technique in order to obtain a new set of passage that is more relevant to the query. Next subsection describes more on the query expansion technique.

[2] http://lucene.apache.org

3.4 Query Expansion

Query expansion is a process of adding a number of words that are relevant to a query, in order for the query to be more clear and unambiguous [4]. There are two types of query expansion methods that are implemented in this system, using PRF and WordNet [2].

Query expansion with Pseudo Relevance Feedback (PRF) .

This technique is based on the n-top passage that obtained using the original query prior to expansion. Each unique word from those passages will be sorted (ranked) based on its TF-IDF value [5]. Every word that has a TF-IDF value higher than the threshold will be acquired as an additional word to expand the initial query. Based on experiments that have been conducted on a number of possible threshold values and their influence on the final score, it is determined that the threshold used in this study was 9.5. TF-IDF method was chosen because it is believed to determine the words that have a strong connection with the initial passage obtained, so that with the addition of the words in the query is expected to obtain correctly relevant passages.

Query expansion with WordNet

WordNet is an online lexical database system, where every noun, verb, adjectives and adverbs will be grouped in a set of synonyms or often called synsets (Synonym sets) [6]. This system uses WordNet to find additional words in queries, which have a semantic relationship, such as hypernym, hyponym, similar adjective and troponym.

3.5 Answer Selection

To determine the answer to a question, it uses a number of feature to assess all the answer choices given (there are 5 pieces of possible answers to a question). When more than one feature is used, the score of an answer choice is accumulated from scores foreach feature. We select an option as the answer of the question which is has the highest final score. Contribution of each feature to score the answer choices vary depending on the weight of the feature. Here is the formula to assign a score to an option C_a.

$$score(C_a) = \sum_{j=1}^{n}(f_i \times w_i) \qquad (3)$$

$f_i$is score of C_a based on i-th feature,w_i is weight of i-th feature,n is number of features to assessC_a.

The system uses a threshold value in selecting the answer, if the difference between answer with the highest score and the score of the other four answer falls below the threshold, then it produces no-answer for the corresponding question.

Answer Selection for Factoid Question.

Finding answers of factoid questions only consisting of a single NE analysis feature which composing into four phases of assessment, which are:

1. Assessment using NE
2. Expansion of the candidate answers for questions of type DATE
3. Assessment using snippet
4. Assessment using average distance weight features and cosine similarity

Initially, this conducts the first assessment phase. If the answer has been found at this stage, the process of finding the answer to a question will be stopped, and three other phases will not be executed. However, if there is finding no answer, then the system carries out the next phase until the fourth phase. When there is still finding no answer after four phases, then the system produce no-answer for the corresponding question.

Assessment using NE.

The process of assessment procedures using NE as follows:

1. Take a set of passage that is relevant to the query
2. Check each passage, whether it has NE or not. When itis find NE, check the type of NE, whether it is similar to the NE of the answer choice. When the NE type of passage is similar tothe type of answer choice, then that NE will be selected to be recorded and given a score.
3. The score of each answer choice is accumulated.
4. The system chooses the answer choice that has the highest score.

Expansion of The Candidate Answers for Questions of Type "DATE".

For the questions of type "DATE", the procedure as follows:

1. If candidates containing the word "later" (eg "9 years later"), but all the answer options that exist in the form of year (eg "1985" or "1997"), then the candidates will be processed first in order to have the same form.
2. If the candidates is in the form of year, while the answer choices are in the range of years(such as "1950s" and the "beginning of 1900") then the candidates will be processed first in order to obtain the proper range.

Assessment Using Snippet.

The system carries out this phase using the search engine Google to acquire additional knowledge. Steps of this phase as follows:

1. Each answer choice is inserted as a query to Google, and the search results that associated with those answer choices will be retrieved. Five pieces of text snippets from the top search resultsare acquired to be an additional knowledge in the form of new documents related to the answer choices. These documents will be indexed with Lucene.
2. Perform step 1 again using keyword from question as the query.

3. Score for an answer choice $C_i$is obtained by accumulating the value of similarity between the questions and snippet belongs C_i, and similarity values between C_i and snippet belongs questions.

$$Score(C_i) = S_1 + S_2 \quad (4)$$

S_1 = Similarity values between the questions and snippet that belongs to C_i.
S_2 = Similarity values between C_i and snippet that belongs to questions.
4. The system chooses the answer choice that has the highest Score.

Assessment Using Average Distance Weight Features and Cosine Similarity.

This phase uses a combination of average distance weight (ADW) and cosine similarity features. Each of these features will file his own score for each answer choice. So for each answer choice, scores of ADW and cosine similarity features will be added to the original score, thus forming a new score. The answer choice having the highest total score is selected as the answer for the question.

Answer Selection forNon-Factoid Question

For non-factoid type questions, selection answers process based on the following eight heuristic features. These features are given a weight based on its ability to determine the correct answer.

Cosine Similarity.

This feature calculates the similarity between two documents based on the weight of each word contained in both documents. This feature implements the concept of vector space model (VSM), where the two documents will be represented as a 2-dimensional vector n (n is the number of unique words), with its components is the weight of each word [4]. The similarity between the two documents will be calculated using the cosine similarity algorithm [2].

The procedure of this feature as follows:

1. Taken 6 sets of passage (assumed as document), whereas one of them is relevant to the question, namely P, and the others are relevant to each choice answers, namely A,B,C,D and E.
2. Each document is represented as a vector, where the components are the weights of each word. For this study, the weight of each word is obtained by taking the value of term frequency using built-in functions provided by Lucene.
3. Finding the cosine value of the angle between two documents. Compute the similarity between document related to the question (P) and documents associated with each choice answers (A, B, C, D and E). The system chooses A, B, C, D or E which having the highest similarity score to P.

Average Distance Weight.

The ADW feature will calculate the smallest distance between each word in answer choice and each word in questions in a set of passage. Steps for the implementation of the ADW feature as follows:

1. Take a set of passage that is relevant to the first answer choice, then combine that into a paragraph.
2. Save all positions of each word in the question and positions of each word in the answer choice that is found in paragraph.
3. Attach each word in the answer choice with each word in the questions, then compute the smallest distance between the positions of both the paired words.
4. Having obtained the smallest distance for each pair of words, add all the distances and divide by the number of word pairs are formed. This outcome is a score for the first answer choice .
5. Do again steps 1-5 for the other four answer choices, in order to obtain a score for each answer choices.

Word Matching.

This feature will calculate the score of each answer choice, based on the number of occurrences of its words on the set of the relevant passage (examined at each passage). The more the words found in the passages, or the more passage that containing the words is found, and then the score for the answer choice will be higher. Likewise, if the words of the answer choices found in the top ranked passage, then the score for this answer choice will also be higher.

Cluster Matching.

In this feature, each answer choice will be given a score based on the value of similarity/fitness between a set of passage that relevant to the answer choice, with a set of passage that relevant to the question. This feature is proposed under the assumption that the answer of the question, of course located adjacent to the questions in a same set passage, so that both set of passage are considered relatively similar.

Question Choice Tilling.

This feature will utilize the maximum number of words in the answer choice that can appear in a sequence in the sentence of passages, which is relevant to the question. The order of appearance of each word from an answer choice will be considered on a sentence in the top 3 passages. For example, for the question "How are people infected by HIV?", If the answer is option "through sexual intercourse", hence the word "through" will have the first position, “sexual" has a second position, and "intercourse" has the third position. When in a sentence in a passage only found the words "sexual" and "intercourse" appearing in sequence, in which the word "through" instead appear after those two words, then the score for this condition is 2. Do this for each sentence in the passage, accumulate the score and grab the highest value.

Maximum occurence.

This feature has a similar function with the word matching feature; both features consider the number of occurrences of answer choice words in a set of relevant passage. The difference lies in the process of checking the appearance of the words. In the word matching feature, any word on the answer choice will be checked at the same time on one passage. So that for each passage, the score only be increased by

the number of words in answer choices at the most, even though the words appear repeatedly in the passage. In this maximum feature, the occurrence of feature will be checked for each sentence in a passage. So, each time the words of that answer choice found in a sentence, then the score will increase by 1.

Maximum Similarity.

This feature using similarity functions in Lucene. At first, it takes sets of passage that relevant to each answer choice. Then, it computes the similarity score between an answer choice to every relevant passage (scores ranging from 0-1) using Lucene. The system chooses the answer choice which has the highest similarity score.

Semantic Relation.

This feature works almost similar to the maximum similarity features; both use the Lucene functions to see the similarity between the answer choices with its relevant passages. This feature uses REVERB to obtain the index; meanwhile, the maximum similarity feature uses reading-test. By using REVERB, it obtains lists of phrases relation that related to the reading-test. Those relation phrases will be indexed by Lucene, resulting in a new index for this feature.

3.6 Evaluation

We use c@1 method to evaluate the performance of this system [7]. The concept of this evaluation method assumes that, instead of giving a wrong answer, the question is better left not have an answer, so the number of wrong answers can be reduced by maintaining the number of correct answers. The formula of evaluation method c@1 as follows:

$$c@1 = \frac{1}{n} \times \left(n_r + \frac{n_r \times n_u}{n} \right) \quad (5)$$

- $n_r$is the number of questions that answered correctly by the system
- $n_u$is the number of questions that are not answered by the system
- n is the total number of questions.

Result of experiment on 2013 datasets

The result of the experiment run describes in the Table 1.

Run	Features	C@1									
		Main					Main+auxiliary				
		1	2	3	4	total	1	2	3	4	total
2	A-W-P-M-C-Ma (0.75)	0.33	0.27	0.24	0.2	0.26	0.33	0.39	0.22	0.27	0.3
3	A-W-M-C-Ma (0.5)	0.3	0.25	0.22	0.2	0.24	0.3	0.38	0.21	0.25	0.29
4	A-W-M-C-Ma (0.75)	0.33	0.25	0.22	0.21	0.25	0.33	0.38	0.21	0.28	0.3
5	A-P-M-C-Ma(0.75)	0.36	0.25	0.23	0.15	0.25	0.36	0.38	0.22	0.23	0.3
6	A-W	0.34	0.22	0.25	0.19	0.25	0.34	0.35	0.23	0.25	0.29
7	A-P-W	0.33	0.23	0.26	0.19	0.26	0.33	0.36	0.25	0.25	0.3
8	A-W-M (0.5)	0.33	0.23	0.25	0.16	0.25	0.33	0.36	0.23	0.21	0.29
9	A-W-Ma (0.75)	0.33	0.25	0.24	0.2	0.25	0.33	0.34	0.23	0.26	0.29
10	A-W-Ma (1)	0.33	0.23	0.23	0.19	0.24	0.33	0.33	0.22	0.25	0.28

Descrition of features:
A = Word Matching, Question Choice Tilling, Semantic Relation, Cosine Similarity & Average Distance Weight
W = Wordnet
P = Pseudo Relevance Feedback
M = Maximum Similarity
C = Cluster Matching
Ma = Maximum Occurrence

Description of topics:
1. Alzheimer
2. Music & Society
3. Climate Change
4. AIDS

According to the Table 1, we can see that the system obtain optimal results when using the two query expansion techniques, along with 5 heuristics features, namely Word Matching, Question Choice Tilling, Semantic Relations, Cosine Similarity and ADW. Two proposed query expansion techniques, PRF and WordNet, both give a positive contribution to the system, especially when they are using together. Meanwhile, among the 8 existing features, features that provide a considerable influence on the performance of the system is 5 features that previously mentioned. The difference in the performance of the system when the weight of features is modified, also became evident that the determination of the appropriate weights for each existing feature, will greatly affect the system performance and the final score.

3.7 Conclusions

Based on the problems posed by QA4MRE, this study has developed a question answering system, which will be used to understand the documents and answer multiple-choice questions that related to documents. To increase the number of relevant

passages that is obtained, this study propose two query expansion techniques, namely PRF and WordNet. In the process of finding answers, the system proposed eight heuristic features that will give scores to each answer choice.

According to the analysis result that obtained in the dataset of 2013, the system will give the best performance when using both the query expansion techniques and all heuristic features with predetermined weights. The result of the final score of the overall system is still relatively low. One reason is indicated because the system does not use the background collection as additional knowledge, so knowledge is only obtained from the documents and snippets from Google. In addition, the system performance when dealing with non-factoid type questions are still not good, so the percentage of questions that answered incorrectly is still very high.

3.8 References

1. Peñas, A., et al. (2011). Overview of QA4MRE at CLEF 2011: Question answering for machine reading evaluation. *CLEF*.
2. Manning, C. D., Raghavan, P., & Schütze, H. (2008, July). *Introduction to information retrieval* (1 ed.). Cambridge University Press.
3. Hatcher, E., & Gospodnetić, O. (2005). *Lucene in action.* United States of America: Manning Publication Co.
4. Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval.* Inggris: Addison-Wesley.
5. Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. New York: McGrawHill.
6. Miller, G. A. (1995, November). WordNet: A lexical database for english. *Communications of the ACM, Vol. 38 No. 11.*
7. Peñas, A., & Rodrigo, A. (2011, June). A simple measure to assess non response. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, (hal. 1415-1424). Portland, Oregon, USA.