

Retrieval-based Question Answering for Machine Reading Evaluation

Suzan Verberne

Information Foraging Lab/Centre for Language and Speech Technology
Radboud University Nijmegen
s.verberne@cs.ru.nl

Abstract. The Question Answering for Machine Reading (QA4MRE) task was set up as a reading comprehension test consisting of 120 multiple-choice questions pertaining to twelve target texts (the test documents) grouped in three different topics. Since this is the first year that we participate in the task, we decided to follow a relatively knowledge-poor approach that is mainly based on Information Retrieval (IR) techniques. We participated in the English task only.

In its most basic version, our system takes the question as a query and uses a standard word-based ranking model to retrieve the most relevant fragments from the test document. Then it matches each of the multiple-choice answer candidates against the set of retrieved text fragments to select the answer with the highest summed similarity (again measured using a standard ranking model). We investigated two forms of information expansion to improve over this baseline: (1) statistical expansion of the test document with sentences from the topical background corpus, and (2) expansion of the question with automatically gathered facts from the background corpus.

Our best-performing experimental setting reaches an overall $c@1$ score of 0.37. We found that statistical expansion of the test documents gives very different results for the three topics but overall it gives an improvement over the document-only baseline. We could not gain any improvements from question-to-facts expansion. More experiments are needed to find a good implementation of fact expansion. In the near future, we will follow up on the current work with more experiments related to expansion of questions and documents for the purpose of question answering.

1 Introduction

After a long tradition of tracks at CLEF in which Question Answering (QA) was a task in its own right, in 2010 and 2011 the organization prepared the QA task as an evaluation method for another task: Machine Reading [3]. According to the track guidelines, “Machine Reading can be defined as the automatic understanding of text. One way of evaluating the understanding of a text is to assess a system’s ability to answer a set of questions about it.”

The Question Answering for Machine Reading (QA4MRE) task was set up as a reading comprehension test. We participated in the English task, which consists of 120 multiple-choice questions pertaining to twelve target texts (the test documents). The test documents are grouped by four in three different topics: AIDS, Climate change and Music & Society. For each of the three topics, a separate background corpus of respectively 29,000, 37,000 and 33,000 documents from a diverse range of sources was provided. In the reading test, five answer candidates are given per question. The participating systems have the choice of not answering a question if they are not sure of the selected answer.

Answering the reading test questions may require more information than the information that is explicitly contained in the test documents. Some examples are:

- A seemingly simple question such as “*Who is the founder of the SING campaign?*” (topic AIDS, document 1, question 2) can be answered by the sentence “And this is the name of my

campaign, SING Campaign.” but only in combination with the knowledge that ‘my’ refers to Annie Lennox, the author of the text.

- The answer to the more complex question “*What event caused Annie Lennox to commit herself to the fight against AIDS?*” (topic AIDS, document 1, question 1) has to be distilled from the first 250 words of the text. The answer “Nelson Mandela’s conference to the world press” is a summarization of Annie Lennox’ motivation.
- The question “*In which European cities has Annie Lennox performed?*” (topic AIDS, document 1, question 7) has as answer alternatives several cities where Annie Lennox performed but which are not located in Europe. Thus, this question can only be answered with knowledge of which cities are in Europe.
- The answer to the question “*In Frankie’s jail, how was diamorphine taken?*” is in the fragment “Frankie, at that time, was a heroin addict, and he was in jail. So his choice was either to accept that dirty needle or not to get high.” but it is difficult to recognize because the synonym *diamorphine* is used instead of *heroin* in the question, and the correct answer “by intravenous injection” does not contain the word *needle*.

Considering these difficulties, the QA4MRE task asks for a number of natural language processing techniques including question analysis (answer type determination), named entity recognition, coreference resolution, synonym/hyponym mapping, etc. Since this is the first year that we participate in the task, we decided to follow a relatively knowledge-poor approach that is mainly based on Information Retrieval (IR) techniques.

In its most basic version, our system takes the question as a query and uses a standard word-based ranking model to retrieve the most relevant fragments from the test document. Then it matches each of the multiple-choice answer candidates against the set of retrieved text fragments to select the answer with the highest summed similarity (again measured using a standard ranking model). We investigated two forms of information expansion to improve over this baseline: (1) statistical expansion of the test document with sentences from the topical background corpus, and (2) expansion of the question with automatically gathered facts from the background corpus.

In Section 2, we will explain our approach and the experiments that we carried out. In Section 3 we describe the results that we obtained. Our conclusions are in Section 4.

2 Our approach

We implemented the QA4MRE task as a pipeline process with five modules: (1) data preparation, (2) test document expansion, (3) question-to-facts expansion, (4) text fragment retrieval and (5) answer selection. Of these, modules 2 and 3 are optional. We implemented all modules in Perl. Then we ran four different experiments:

1. no expansion;
2. test document expansion;
3. question-to-facts expansion;
4. test document expansion and question-to-facts expansion.

In all these experiments, modules 1, 4 and 5 are used. Module 2 is active in experiment 2 and 4; module 3 in experiment 3 and 4.

2.1 Data preparation

We prepared the test documents and the background corpus as follows: In the test documents, we replaced the words ‘I’, ‘me’, and ‘my’ by the name of the author, which we simply assumed to be the first name of the document. We split the test documents in target fragments of three sentences each, and the background corpora in single-sentence units. Then we lowercased all texts

and removed punctuation. We created an inverted index of each topic’s background corpus and we removed all words from the index that occur in fewer than three sentences.¹

2.2 Test document expansion

In Section 1, we showed that some of the questions of the reading comprehension task cannot be answered with only the information contained in the corresponding test document; additional information from the background corpus is required to answer these questions. Our document expansion step in experiment 2 and 4 is based on the idea that the required additional information is in some way related to the information that is already present in the test document. We implemented a statistical expansion module that searches the most similar sentences in the topic’s background corpus index for each of the three-sentence fragments (the ‘target fragments’) from the test document.

The similarity between the target fragments and the background sentences is measured using the BM25 ranking function [1]: For each target fragment T , containing keywords t_1, \dots, t_n , the BM25 score of a background sentence S is:

$$score(S, T) = \sum_{i=1}^n ISF(t_i) \cdot \frac{f(t_i, S) \cdot (k_1 + 1)}{f(t_i, S) + k_1 \cdot (1 - b + b \cdot \frac{|S|}{avgsl})} \quad (1)$$

in which $f(t_i, S)$ is t_i ’s term frequency in S , $|S|$ is the length of S in words, and $avgsl$ is the average sentence length in the background collection. For the parameters k_1 and b , we used $k_1 = 2.0$ and $b = 0.75$. $ISF(t_i)$ is the Inverse Sentence Frequency of the target term t_i :

$$ISF(t_i) = \log \frac{N - n(t_i) + 0.5}{n(t_i) + 0.5} \quad (2)$$

where N is the total number of sentences in the collection, and $n(t_i)$ is the number of sentences containing t_i .

Our document expansion module expands each target fragment with the ten highest-scoring sentences from the background collection that are at least four words long.² Table 1 shows an example target fragment with the ten most related sentences that are retrieved from the background corpus. The result is a set of expanded target fragments ET for each test document.

2.3 Question-to-facts expansion

In the previous subsection we described a statistical method for expanding the test document with information from the background corpus. A second type of expansion that we investigated, is expanding the question itself with structured information from the corpus. The rationale behind this is that each question contains names and domain-related terms about which the background corpus contains factual information. This factual information might help for answering the question. For example, the corpus contains facts about needle exchange programs, e.g. “needle exchange programs, which have proved to reduce the spread of blood-borne viruses” (document AIDS_EN_10645.txt). This information can help answering the question “Why did a former UK Prime Minister start a needle exchange program?” (answer: “in order to prevent the spread of HIV”)

In the remainder of the current subsection, we explain how we extracted facts from the corpus; in the next subsection we explain how we used the question-to-facts expansion in the QA process.

We implemented a fact extractor for English that is based on dependency relations, similar to the method proposed in [6]. Each fact has the general form `subject|verb|object|complements` or

¹ Removing words from the index that have a sentence frequency below three also had a practical reason: Perl’s data structures are not memory-efficient enough to store an inverted index including all hapaxes and near-hapaxes.

² Without the length requirement, BM25 gives us too many ‘sentences’ of only one or two words long.

Table 1. Target fragment expansion example. The first line contains the target fragment: the first three sentences from document 1 on topic ‘AIDS’, lowercased and without punctuation. The next lines contain the most similar sentences from the background corpus on AIDS, together with the IDs of the document they come from and their BM25 score. Note that during matching, the words ‘I’, ‘me’ and ‘my’ have been replaced by the name of the author of the document, ‘Annie Lennox’.

why i am an hiv aids activist im going to share with you the story as to how i have become an hiv/aids campaigner and this is the name of my campaign sing campaign		
Doc-ID	Normalized sentence text	BM25 score
5803	hiv / aids awareness and care campaign	25.8
3135	im going to sing dixie until she cries	24.3
26832	new campaign focuses on hiv/aids new campaign focuses on hiv/aids by janet kornblum usa today	23.6
2878	reports from the holocaust the story of an aids activist	23.4
11548	an environmental campaigner and aids activist hu has become one of the most outspoken critics of chinas human rights violations	22.9
13149	so im going to sing this to you now so you know how it will sound at the funeral	22.7
2404	the campaign is going to focus on educating people about the threat that aids poses	22.4
9497	health activist tim bannet of the world aids campaign shares richters view	22.3
26745	aids activist welcomed the decline but questioned whether the governments anti-aids campaign was a success	22.3
11873	annie lennox is the founder of the sing campaign which works with comic relief to tackle hiv	22.1

subject|verb|predicate|complements, in which verbs and nouns are lemmatized. Here are three examples of facts automatically extracted from the background collection on AIDS:

“needle exchange program | reduce | the spread | ”
 “Queen II Elizabeth | honor | Annie Lennox | ”
 “the future prime minister | write | romance novel | possibly

We used the AEGIR parser [5] to generate the dependency relations needed for the fact extractor. AEGIR was developed for applications in Information Retrieval rather than in Linguistics. For that reason, its dependency model expresses the *aboutness* [4] of a sentence rather than describing its complete syntactic structure. AEGIR’s output representation is comparable to the Stanford typed dependencies representation [2], in the sense that it generates a set of binary relations between words for an input sentence, thereby converting some function words (such as prepositions) to relations. In addition to that, AEGIR performs a number of normalizing transformations, such as passive-to-active transformation. For example, the clause “an inflammatory reaction, caused by the bowel tissue” leads to the same analysis as “the bowel tissue causes an inflammatory reaction”:

“the bowel tissue | cause | an inflammatory reaction | ”

We parsed the background corpora with AEGIR (version 1.8.2) and extracted the facts using a Perl script. We performed some automatic postprocessing to filter out ‘empty’ facts of which the subject or the object is a pronoun³ and facts that come from metadata instead of actual document content (e.g. containing the word ‘Wikipedia’).

We considered several options for expanding the input questions with facts from the background corpus, and we decided on a rather basic approach: We indexed the facts twice, by subject (first field) and by object (third field). Then, while processing the question Q , if any substring of the question acts as subject or object in an indexed fact F_i , this fact is added to the question expansion

³ Of course, better would be to replace these anaphors by their antecedent but we did not add an anaphora resolution step in this version of our method.

set EF_Q . We set a threshold on the frequency of the fact in the background corpus that has to be exceeded for the fact to be included in EF_Q .

For example, the question “Why did a former UK Prime Minister start a needle exchange program?” (topic AIDS, document 3, question 6) contains three substrings that act as subject or object in facts from the background corpus: ‘prime minister’, ‘needle exchange’ and ‘needle exchange program’. Example background facts containing these terms are:

the country | have | prime minister | now
 an order | permit | needle exchange | in three city
 needle exchange program | decrease | the spread |

EF_Q , the set of facts generated for a question, can be very large if frequent substrings (e.g. ‘problem’, ‘people’) from the question act as subject or object in facts. Therefore, we set a maximum sentence frequency of 2,000 and a minimum length of 7 characters on each substring. If a substring is more frequent or shorter, then no facts are generated for this substring.

2.4 Text fragment retrieval

The retrieval module retrieves the most relevant target fragments for each question. If set to active, the document expansion module has before expanded each three-sentence fragment in the test document with ten sentences from the background corpus.

For measuring the similarity between the question and an (expanded) target $score(Q, ET_i)$ we again use the BM25 ranking function. For each question, up to ten expanded fragments are retrieved for which the BM25 score is higher than a fixed threshold. The result is a set of (expanded) target fragments that are relevant to the question Q : ET_Q .

If the question-to-facts expansion module is active (in experiment 3 and 4), then we add the selected facts EF_Q to ET_Q . For example, in experiment 2, two (non-expanded) fragments from the test document are retrieved for the question “Why did a former UK Prime Minister start a needle exchange program?” (topic AIDS, document 3, question 6). In addition to this, the question-to-facts expansion module selected 12 facts related to the question terms ‘prime minister’, ‘needle exchange’ and ‘needle exchange program’. The two fragments and the twelve facts together form the question expansion set E_Q .

2.5 Answer selection

The last step of our pipeline is to select one of the multiple-choice answer options from the reading test. We implemented a simple voting system, again based on BM25, for selecting the most likely answer given E_Q ⁴: We treat each answer candidate as a query, and calculate its BM25 score for each of the elements in $E_{Q,i}$. Then we sum this score over all elements to get a score for each answer:

$$score(A_j, E_Q) = \sum_{i=1}^n score(A, E_{Q,i}) \quad (3)$$

where n is the number of (expanded) target fragments and facts for Q .

We set a threshold on the answer selection score: If the $score(A_j, Q)$ for none of the answer candidates A_j exceeds this threshold, then the question is not answered.

3 Results

As introduced in the previous section, we submitted four experiments:

⁴ Recall that E_Q contains text fragments from the test document; depending on which experiment we run these are statistically expanded to ET_Q , and/or extended with the set of facts EF_Q .

1. no expansion;
2. test document expansion;
3. question-to-facts expansion;
4. test document expansion and question-to-facts expansion.

3.1 The effects of selective thresholds

In Section 2, we mentioned that we used thresholds at several points in our method.

In the **question-to-facts expansion module**, we applied a threshold to the frequency of a fact in the corpus for it to be selected for a question’s expansion set. We noticed that there is quite some noise among the facts that only occur once in the corpus. Therefore, we require that the absolute fact frequency be ≥ 2 . Table 2 shows the number of questions that are expanded with at least one fact from the background corpus, and the average number of facts per expanded question per topic.

Table 2. Statistics on question-to-facts expansion. The reading test consists of 40 questions per topic.

	AIDS	Climate change	Music & Society
No. of expanded questions	2	11	4
Avg. no. of facts per expanded question	14.5	99.7	3.8

Considering these statistics, it is clear that more work is needed on sensible expansion of questions with facts. The low numbers of questions that get expanded with facts, especially for the AIDS and Music & Society topics, suggests that our method for selecting substrings for expansion is too rigid. We will look into this in more detail before following up with additional experiments. Moreover, the Climate Change topic has an extremely high number of facts per expanded question. This is because the question substrings ‘carbon dioxide’ and ‘climate change’ are both very frequent in the Climate Change corpus, and generate large amounts of facts, e.g.

fossil fuel | generate | carbon dioxide
the requirement | reduce | carbon dioxide
the causal | link | carbon dioxide

The first of these facts may be relevant for the question “How could the reduction of carbon dioxide emissions be promoted?”, but the second fact is not informative, and the third has incorrectly been generated due to parsing errors. An interesting challenge for future work is to estimate the informativeness of facts for question expansion.

Secondly, we applied a threshold on the BM25 score in the **text fragment retrieval module**. Up to ten fragments were retrieved per question, but if fewer than ten exceed the threshold then only those are retrieved that do. We used a threshold of $score(Q, ET_i) = 2$. On average, 2.0 fragments were retrieved per question in experiment 1 and 3 (without document expansion); 4.9 fragments were retrieved per question in experiments 2 and 4 (when the fragments have been expanded with information from the corpus).

Finally, we applied a threshold on the BM25 score in the **answer selection module**, which we use to decide whether or not our system answers the question: if it is not certain enough of the highest-scoring answer, then it decides not to answer the question. We used a threshold of $score(A_j, E_Q) = 1$. Table 3 shows the number of unanswered questions per experimental setting per topic.

3.2 Evaluation of the runs

The main evaluation measure used for the QA4MRE runs is $c@1$ [7]. This measure not only counts the number of correctly answered questions, but also takes into account the number of questions that are not answered by a system:

Table 3. Number of unanswered questions, per setting per topic. The reading test consists of 40 questions per topic.

Setting	AIDS	Climate change	Music & Society
1. no expansion	12	12	8
2. test document expansion	5	0	2
3. question-to-facts expansion	12	12	8
4. test document expansion and question-to-facts expansion	5	0	2

Table 4. The four experimental settings evaluated with $c@1$, overall and per topic. Per topic, the best scoring setting(s) is/are printed in bold face.

Setting	Overall	AIDS	Climate change	Music & Society
1. no expansion	0.34	0.26	0.29	0.45
2. test document expansion	0.37	0.28	0.45	0.37
3. question-to-facts expansion	0.33	0.23	0.29	0.45
4. test document expansion and question-to-facts expansion	0.35	0.28	0.40	0.37

$$C@1 = (nr + nu * (nr/n))/n \quad (4)$$

where nr is the number of correctly answered questions, nu is the number of unanswered questions and n is the total number of questions.

Table 4 shows the results for our runs in terms of $c@1$.

We make several observations from these results. First, experimental setting 2 gives the best results over all topics. Music & Society is an exception to this trend; here the settings without expansion of the test documents outperform the settings *with* document expansion. We had a look at the expanded fragments from the source documents and we suspect that the specificity of the test documents (e.g. one specific music video) turns the most similar sentences from the background corpus (which are about other music videos) irrelevant for answering the questions. For the other two topics, expansion of the test document with sentences from the background corpus gives an improvement. This improvement is very large for the Climate Change topic (from $c@1 = 0.29$ to $c@1 = 0.45$).

Our second observation is that question-to-facts expansion does not improve the performance of our system on the reading tests. As we explained in Section 3.1, we expect that there is a lot to gain from sensible expansion of questions with facts: First, in the quality of the facts generated from the background corpus (which depends on the quality of the dependency parser), second in the selection of informative facts to be included in the expansion set, and third in the selection of the substrings (terms) that are expanded to facts.

4 Conclusions

The QA4MRE task is a reading comprehension task: the goal is to answer a number of multiple-choice questions pertaining to a test document, possibly with use of a background corpus on the topic of the document.

We implemented a relatively knowledge-poor approach to the reading comprehension task. Our approach is based on retrieval techniques in two steps: retrieval of relevant fragments from the test document for the input question, and matching of the multiple-choice answer candidates against the retrieved fragments in order to select the most likely answer.

In a series of experiments, we investigated two forms of information expansion: (1) statistical expansion of the test document with sentences from the topical background corpus, and (2) expansion of the question with automatically gathered facts from the background corpus.

We found that statistical expansion of the test documents gives very different results for the three topics but overall it gives an improvement over the document-only baseline. We could not

gain any improvements from question-to-facts expansion. More experiments are needed to find a good implementation of fact expansion.

In the near future, we will follow up on the current work with more experiments related to expansion of questions and documents for the purpose of question answering.

References

1. S. Büttcher, C.L.A. Clarke, and B. Lushman. Term Proximity Scoring for Ad-hoc Retrieval on Very Large Text Collections. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 621–622. ACM, 2006.
2. M.C. De Marneffe and C.D. Manning. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics, 2008.
3. O. Etzioni, M. Banko, and M.J. Cafarella. Machine reading. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, pages 1517–1519. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.
4. C.H.A. Koster, J.G. Beney, S. Verberne, and M. Vogel. Phrase-Based Document Categorization. *Current Challenges in Patent Information Retrieval*, pages 263–286, 2011.
5. Nelleke Oostdijk, Suzan Verberne, and Cornelis H.A. Koster. Constructing a broad coverage lexicon for text mining in the patent domain. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*. European Language Resources Association (ELRA), 2010.
6. A. Peñas and E. Hovy. Semantic enrichment of text with background knowledge. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 15–23. Association for Computational Linguistics, 2010.
7. A. Penas, A. Rodrigo, and J. del Rosal. A Simple Measure to Assess Non-response. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1415–1424. Association for Computational Linguistics, 2011.