Combining text mining techniques for QA4MRE 2013

Guilherme de Oliveira da Costa Marques[1] and Mathias Verbeke[2]

[1]Federal University of São Carlos (UFSCar), campus Sorocaba, Brazil
[2]Department of Computer Science, KU Leuven, Belgium

Abstract. This paper describes a lexical system developed for the main task of Question Answering for Machine Reading Evaluation 2013 (QA4MRE). The presented system executes the preprocessing of test documents, and generates hypotheses consisting of the question text combined with text from possible answers for the question. The hypotheses are compared to sentences from the text by the means of a set similarity measure. The k best similarity scores obtained by each hypothesis are averaged as ranking score for the hypothesis. Two variations of the developed system were utilized, one of them employing coreference detection and resolution techniques in order to take advantage of the discourse structure on the question answering process. The results generated by the systems in QA4MRE 2013 edition are presented and analyzed. The presented system should serve as a solid base for the development of a semantic approach on the task.

1 Introduction

The QA4MRE competition [1] focuses on the Question Answering field of Machine Reading. Adopting the form of several tests spread over a few themes, it aims at evaluating a system's Natural Language Understanding capabilities by the means of multiple choice questions.

The main task of the competition is currently composed by four topics: "AIDS", "Climate Change", "Music and Society" and "Alzheimer's". A background collection of texts is provided for all topics. This collection attempts to encompass all specific domain knowledge of the topic.

The 2013 edition consists of 16 reading tests, 4 on each topic. Each reading test presents a text document followed by 15 to 20 questions about it. Those are multiple choice questions with 5 alternatives, the last one being "None of the above".

Questions are distributed over different degrees of complexity as to the knowledge and inference required to devise the correct answer. The simplest ones have both the question fact and the answer appearing directly in the same sentence of the text. Others have the question fact and the answer appearing in distinct sentences. Some questions require background knowledge or inference, and some may require the use of both.

Question: *What caused an improvement in sound quality in 1950?*
Alternatives:

1. the introduction of soundtrack recording on 35 mm magnetic tape
2. the use of an optical soundtrack
3. the adoption of a quadraphonic sountrack
4. the specialisation in silent films
5. none of the above

Table 1: Example question

An example question is presented in table 1. For this question, alternative 1 should be identified as correct.

The system described in this work is based on a *text mining* baseline system [2]. It was developed and tested with data from QA4MRE 2012 edition. Several parameters and system variations were tested, which are described more thoroughly in [3]. We do not employ the background collection in the current system.

2 Methodology

Two different systems were employed for the competition: a main system was elaborated, and employed both as standalone and as a base for a variation including *coreference resolution*. The structure of both systems is presented in figure 1.

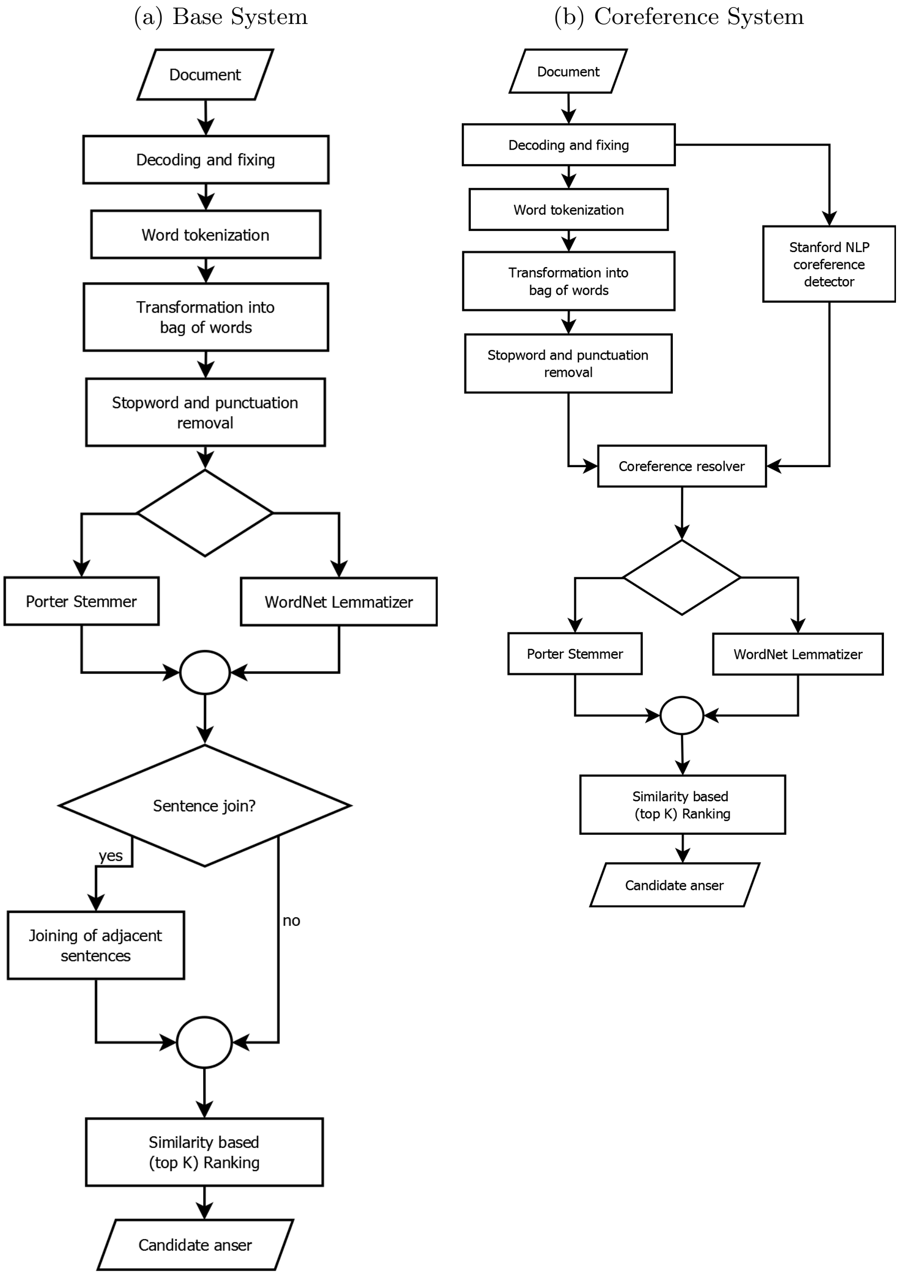
Section 2.1 presents the text preprocessing employed on the test documents. Section 2.2 explains the procedure responsible for ranking the alternatives, while section 2.3 discusses the different techniques employed in cases where a tie occurs in the ranking. While all the previously mentioned characteristics are common to the two systems, section 2.4 explains the additions only present in the *coreference variation*.

2.1 Preprocessing

Preprocessing of the test documents proceeds according to the following sequence:

1. **Unicode decoding:** treats special unicode characters, generating ASCII-safe strings.
2. **Text fixing:** corrects some of the formatting issues present in the 2012 test set, through the use of regular expressions. Although similar issues were not detected with 2013's test set, the procedure was maintained for safety and compatibility reasons.
3. **Sentence tokenization:** splits sentences from text into separate strings.
4. **Word tokenzation:** segments individual words and punctuation signs from text strings.

Fig. 1: System structure



(a) Base System

(b) Coreference System

Document
Decoding and fixing
Word tokenization
Transformation into bag of words
Stopword and punctuation removal
Porter Stemmer
WordNet Lemmatizer
Sentence join?
yes
no
Joining of adjacent sentences
Similarity based (top K) Ranking
Candidate anser
Document
Decoding and fixing
Word tokenization
Transformation into bag of words
Stopword and punctuation removal
Stanford NLP coreference detector
Coreference resolver
Porter Stemmer
WordNet Lemmatizer
Similarity based (top K) Ranking
Candidate anser

5. **Bag of words model:** sentences are represented as sets of word strings.
6. **Stopword and punctuation removal:** punctuation signs and words classified as stopwords according to the English stopword corpus from NLTK [4] are removed from the sentences.
7. **Stemming or Lemmatization:** words are converted into word *stems* (by NLTK's Porter stemmer) or into *lemmas* (by NLTK's WordNet-based lemmatizer).
8. **Joining of adjacent sentences (optional):** in an attempt to perform a primitive form of discourse analysis, a procedure where the sets of words from adjacent sentences are joined prior to ranking was tested. Results generated by this strategy on QA4MRE'12 test set presented an overall improvement in accuracy, so this procedure was maintained for the base system.

With respect to item 7, the choice between *stemming* and *lemmatization* took into consideration the accuracy observed on 2012 test set. For the base system, *lemmatization* yielded marginally better results, while the *coreference* system presented improved results when *stemming* was employed.

2.2 Ranking procedure

Ranking is performed by computing the similarity between sentences and hypotheses, as presented in Algorithm 1. The hypotheses are generated by joining *question text* and the text from alternatives:

$$H_i = Q \cup A_i$$

Algorithm 1 Ranking Procedure

```
$ranks \leftarrow list()$
for all hypothesis $hyp$ in hypotheses list do
    $similarities \leftarrow list()$
    for all sentence $sent$ in document do
        $sim \leftarrow similarity(sent, hyp)$
        $similarities.append(sim)$
    end for
    $average \leftarrow 0$
    for all top k values $sim$ in $similarities$ do
        $average \leftarrow average + sim$
    end for
    $average \leftarrow average/k$
    $ranks.append(average)$
end for
$selected \leftarrow indexOf(maximum(ranks))$
```

The similarity metric employed during the ranking procedure is the MASI similarity [5], calculated by the formula below:

$$masi\_sim(set_1, set_2) = \frac{|set_1 \cap set_2|}{max(|set_1|, |set_2|)}$$

According to the similarity scores calculated between hypotheses and text sentences, each hypothesis receives an overall ranking score computed as the average of the k best sentence similarities. The candidate answer chosen by the system corresponds to the hypothesis which presents the highest ranking score. The value of k is a parameter provided to the system, and was set to $k = 2$, considering the test results obtained with the QA4MRE'12 test set.

2.3 Handling of ties

In some cases, the ranking procedure results in a tie between two or more hypotheses. To handle those cases, four different strategies were envised:

1. All questions where no candidate answer was found (there was a ranking tie) were answered as "*None of the above*".
2. All questions where no candidate answer was found were left unanswered.
3. If the maximum ranking score between the alternatives is inferior to a certain threshold, the question is answered as "*None of the above*". Otherwise, it is left unanswered.
4. If the maximum ranking score between the alternatives is inferior to a certain threshold, the question is answered as "*None of the above*". Otherwise, one of the tied alternatives is selected at random.

The reasoning behind the threshold value utilized in 3 and 4 is that in questions where the ranking values were lower, there would be a higher chance that none of the alternatives was correct and the question had no answer. In contrast, in questions where the ranking values were higher, it would be more likely that there was a correct answer, but the system was unable to find it. This threshold was empirically set to *0.1*.

2.4 Coreference variation

The system illustrated in Figure 1b includes a *coreference detection* phase, as well as a *coreference resolver*. This addition intends to take advantage of discourse analysis, allowing for the resolution af anaphoric pronouns, as well as other types of coreferences.

Coreference detection is performed by Stanford's NLP suite [6], which outputs an XML file containing information on the detected coreferences present in the processed text. Information extracted from this XML file is employed by the coreference resolver in the following way:

1. The representative *noun phrase* is located.
2. Words from the representative reference are included in a word set.

3. Sentences where other references to the same entity appear are located.
4. The word set from the representative reference is joined into the sentences that refer to the same entity.

This simple strategy presents good results with regards to the resolution of referential pronouns, in the context of a hypothesis ranking computed through set similarity: since the words from the representative reference are included in the word set of referencing sentences, this has a positive impact on the similarity between those sentences and hypotheses that mention the same entity.

3 Results

Eight distinct runs were submitted to the competition, where the two presented systems were paired with each one of the four tie strategies described in section 2.3. General results from each run are presented in table 2. In this table, performance is measured according to two metrics: accuracy and the c@1 measure [7]. C@1 is the main performance measure employed in the competition, and is calculated as follows:

$$c@1 = \frac{(n_R + n_U * \frac{n_R}{n})}{n}$$

where
n_R = number of correctly answered questions
n_U = number of unanswered questions
n = total number of questions

The competition consisted of a total of 240 main questions, of which 44 required inference in the answering process. Those inference-demanding questions had simpler duplicates where the question was phrased in a way the inference was no longer required. The "*c@1 main*" accuracy only takes the 240 main questions into consideration, and "*c@1 all*" is calculated over all 284 questions.

Algorithm			Main Qs.		All Qs.	
System	Tie	Run	Accur.	c@1	Accur.	c@1
Base + Join	1	02	0.28	0.28	0.34	0.34
	2	03	0.23	0.26	0.29	0.33
	3	04	0.27	0.28	0.33	0.35
	4	05	0.28	0.28	0.35	0.35
Coreference	1	06	0.30	0.30	0.33	0.33
	2	07	0.22	0.26	0.26	0.30
	3	08	0.26	0.29	0.29	0.32
	4	09	0.28	0.28	0.31	0.31

Table 2: General results

Run number 6 presented the best general results, with a c@1 measure of 0.30 on the main questions. The considerable increase in the c@1 metric between the main set of questions and the complete set reinforces the weakness of the employed systems with inference demanding questions.

Algorithm			topic 1		topic 2		topic 3		topic 4	
System	Tie	Run	Median	Average	Median	Average	Median	Average	Median	Average
Base + Join	1	02	0.30	0.33	0.21	0.18	0.22	0.25	0.22	0.21
	2	03	0.26	0.29	0.18	0.16	0.21	0.23	0.19	0.20
	3	04	0.31	0.33	0.19	0.17	0.22	0.25	0.23	0.22
	4	05	0.33	0.33	0.18	0.17	0.22	0.25	0.22	0.22
Coreference	1	06	0.37	0.37	0.18	0.19	0.26	0.27	0.17	0.21
	2	07	0.23	0.26	0.20	0.20	0.23	0.23	0.17	0.18
	3	08	0.28	0.31	0.19	0.21	0.27	0.26	0.17	0.20
	4	09	0.27	0.28	0.18	0.19	0.28	0.26	0.19	0.21

Table 3: C@1 per topic

Table 3 lists the c@1 results per topic, for the main set of questions. From the analysis of this table, a considerable variation in performance between topics can be noticed: the system presented the best results in topic 1, followed by topic 3; topics 2 and 4 have inferior performance. Another remarkable fact is that while the best results from the coreference system are superior to the results from the base system in topic 1, the differences in the results on other topics are negligible.

4 Conclusions

The presented methodology is entirely based on lexical similarity. Possible directions for improvement are the inclusion of techniques for proper handling of questions involving negation (added to 2013 main task, but not present in 2012). The system could also benefit from a weighted similarity measure that would prioritize words according to importance.

Although there is still room for improvement while maintaining the lexical character of the system, we believe that the ideal focus of future work would be on establishing a system able to deal with semantic relations through the development of strategies aiming at textual and logic inference. We also consider of crucial importance the development of techniques for knowledge base construction, which can perform the extraction of domain-specific knowledge from the background collection.

5 Acknowledgements

Guilherme de Oliveira da Costa Marques is funded by Brazil's National Counsel of Technological and Scientific Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq). Mathias Verbeke is funded by the Research Foundation Flanders (FWO-project G.0478.10 - Statistical Relational Learning of Natural Language).

References

1. CELCT: Question Answering for Machine Reading Evaluation http://celct.fbk.eu/QA4MRE/.
2. Verbeke, M., Davis, J.: A text mining approach as baseline for QA4MRE'12. In: CLEF (Online Working Notes/Labs/Workshop). (2012) http://www.clef-initiative.eu/documents/71612/234cb84c-03a3-45c3-8844-9b1ca448d976.
3. de Oliveira da Costa Marques, G.: Combining text mining techniques for question answering. Master's thesis, KU Leuven (2013)
4. Bird, S., Klein, E., Loper, E.: Natural language processing with Python. O'Reilly (2009)
5. Passonneau, R.: Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC). (2006) 831–836
6. Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics (2013) 1–54
7. Peñas, A., Forner, P., Sutcliffe, R.F.E., Rodrigo, Á., Forascu, C., Alegria, I., Giampiccolo, D., Moreau, N., Osenova, P.: Overview of ResPubliQA 2009: Question answering evaluation over european legislation. In: CLEF. (2009) 174–196 http://dx.doi.org/10.1007/978-3-642-15754-7_21.