Using Anaphora resolution in a Question Answering system for Machine Reading Evaluation

Adrian Iftene[1], Alex Moruz[1, 2], Eugen Ignat[1]

[1] UAIC: Faculty of Computer Science, "Alexandru Ioan Cuza" University, Romania
[2] Institute of Computer Science, Romanian Academy Iasi Branch
{adiftene, amoruz, eugen.ignat}@info.uaic.ro

Abstract. This paper describes UAIC[1]'s Question Answering for Machine Reading Evaluation systems participating in the QA4MRE 2013 evaluation task. We submitted two types of runs, both type of runs based on our system from 2012 edition of QA4MRE, and both used anaphora resolution system. Differences come from the fact the textual entailment component was used or not. The results offered by organizer showed that runs based on textual entailment component were better.

Keywords: Question Answering for Machine Reading Evaluation, Information Retrieval, Textual Entailment, Anaphora Resolution

1 Introduction

As in the 2012 campaign, the Question Answering for Machine Reading Evaluation (QA4MRE[2]) task in 2013 intends to cross-evaluate the ability of systems to read and understand texts[3]. The systems involved in this task must have ability on reading single documents and to identify the correct answer from a set of five multiple choice answers, using different kinds of inferences and previously acquired background knowledge. The background knowledge is based on 2012 collection and it is related to four topics: *AIDS*, *Climate Change*, *Music and Society*, *Alzheimer* [1]. In 2013 was involved 5 different languages (Arabic, Bulgarian, English, Romanian and Spanish), and the test data was the same (parallel translations) and the background knowledge was available to all participants. In comparison with previous editions this year appear two main differences: (1) was inserted questions based on modality and negation aspects, and (2) a portion of questions have no correct answer and the correct option was "none of above".

The system used by our group in 2013 QA4MRE edition is an improved version of the system used in 2012 QA4MRE edition [2]. The system from 2012 was further improved by adapting an anaphora resolution component for the Question Answering module.

---

[1] University "Al. I. Cuza" of Iasi, Romania
[2] QA4MRE: http://celct.fbk.eu/QA4MRE/index.php
[3] QA4MRE Guidelines: http://celct.fbk.eu/QA4MRE/scripts/downloadFile.php?file=/websites/ResPubliQA/resources/guideLinesDoc/main/QA4MRE2013_Guidelines.pdf

The rest of the paper is structured as follows: Section 2 details the general architecture of our Question Answering system for Machine Reading Evaluation and the new textual entailment module, Section 3 presents the results and an error analysis, while the last Section discusses the conclusions.

2 System components

In QA4MRE 2013, UAIC submitted runs only for Romanian. For that, we use the system from the previous edition of QA4MRE 2012 [2], consisting in modules specialized for *test data processing, background knowledge indexing*, *snippet extraction*, *identification of the correct answer* and *textual entailment*. In addition in pre-processing part this year we used the anaphora resolution component.

2.1 The base architecture

In 2013, the Romanian background knowledge was based on version from 2012 and it consisted of a collection of 184,263 documents in text format (32,631 correspond to the *AIDS* topic, 19,190 to *Alzheimer* topic, 63,207 to *Climate Change* topic and 92,268 to *Music and Society* topic). The test data consists in an XML file with 16 test documents (4 documents for each of the four topics), 15 to 20 questions for each document (284 questions in total) and 5 possible answers for each question (1,420 possible answers in total).

The base architecture is similar to the system used for the 2012 edition of the QA4MRE competition, presented in [2]. Thus, after indexing the background collection using Lucene[4] libraries [7], the system processes the test data applying 3 operations: (a) extracting documents from the background knowledge, (b) analyzing the test questions and (c) processing possible answers. If the first step is performed using Lucene indexing of the background collection, for analyzing the question we used our question processing module [2] and the web services available from the Sentimatrix[5] project [5] to eliminate stop words, perform lemmatization and identify the Named Entity in the question. Then, a Lucene query is built. For instance, in the case of the question with q_id = "1":

Ro: *Cu ce scop unii cobai au fost injectați cu gene umane care provoacă Alzheimer*?
En: *For what purpose were some mice injected with human genes that cause Alzheimer's*?

the execution of the above steps has the following results:

- in the first step, the following stop words are eliminated: *cu, ce, unii, au, fost, cu, care* (En: *for, what, were, some, with, that, 's*);

[4] Lucene: http://lucene.apache.org/
[5] Sentimatrix: http://www.sentimatrix.eu/

- in the next step, lemmas for the words *injectaţi, gene, umane, provoacă* (En: *injected, genes, human, cause*) are identified;
- in the third step, *Alzheimer* is identified as a Named Entity;
- in the last step, the Lucene query is build: " *scop cobai (injectaţi^2 injecta) (gene^2 geană) (umane^2 uman) (provoacă^2 provoca) Alzheimer^3*".

From the above Lucene query, one can notice that we consider named entities to be of most relevance (hence receiving a boost of 3, expressed as using the ^ operator), while the inflected form of the words existing in the question receive a lower boost value (2 in the example above).

Another module analyzes the possible answers types and features, using the ontology presented in [6], more specifically the relations between regions and cities and the relations between cities and countries, in order to eliminate the answers with low probability to be the required answer. For instance, for the question with q_id="14":

Ro: *Ce ţară este liderul REDD în America?*
En: *Which country is the leader of REDD in America?*,

we eliminate from the list of possible answers the answers with non-American states.

As presented in [2], the index of background knowledge is queried, and all retrieved documents are placed in separate indexes. The results of this step are 284 separate indexes for every question from the initial test data. Then every index is searched for every answer, and a list of documents with Lucene relevance scores are returned, where *Score(d, a)* is the relevance score for document *d* when we search with the Lucene query associated to the answer *a*.

In 2013 we submitted two types of runs: (1) first without textual entailment module, and (2) second using textual entailment module. In the first case, after above steps a normalized value is computed for all answers associated to a question, and the answer with the highest value is selected as the most probable answer. In the second case, we perform the following three steps (i) we build a pattern with variables for every question according to the question type; (ii) using a pattern and all possible answers, we build a set with 5 hypotheses for each of the questions; (iii) we assign to the document tag from the initial XML file the role of text T and we run the TE system for all obtained pairs [3, 4].

2.2 Anaphora resolution

For the QA4MRE 2013 besides the improved system from QA4MRE 2012, we added anaphora resolution based run. Anaphora resolution is defined as the process of resolving an anaphoric expression to the expression it refers to [8]. The UAIC developed tool that handles anaphora resolution is called RARE (Robust Anaphora Resolution Engine) and uses the work done in [9], where the process uses three layers (Figure 1):

- The text layer, which contains referential expressions(RE's) as they appear in the discourse;
- An intermediate layer (projection layer) that contains any specific information that can be extracted from the corresponding referential expressions;
- A semantic layer that contains the descriptions of the discourse entities (DE). Here the information contributed by chains of referential expressions is accumulated.

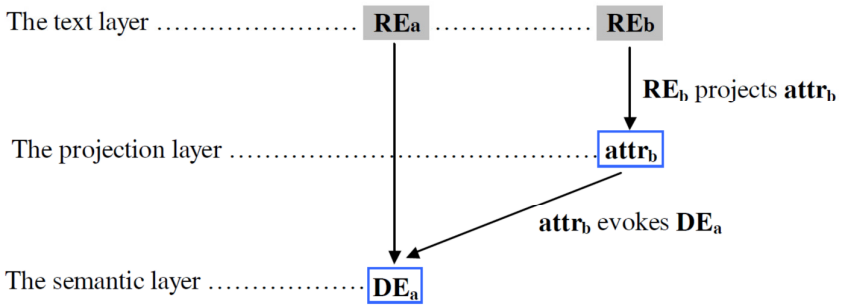

Figure 1: Three layers structure [9]

The core of the RARE system is language independent, yet to be able to localize it to one language it needs specific resources. These specific resources are:

- **Constrains** – a file that contains the rules which will match the conditions between anaphor and antecedent;
- **Stopwords** – a file that contains a common list of stop-words;
- **Tagset** – a file that contains the mapping from the tagset used in the input file to a simplified tagset used by the system;
- **Window** – a file that contains the length of the window where the antecedent should be looked by the system; the length is in tokens.

The process of anaphora resolution runs in the following chain: First the input is read from left to right; when a new noun phrase is found, a new referential expression (RE) is created which contains the morphologic, syntactic and semantic features; all these features are then tested by the rules defined in the constraints file and it is decided where this new RE defines a new discourse entity or it refers a before mentioned one, and finds which one. The system outputs chains of co-referential expressions. Each such chain is characterized by a feature structure that sums up all the features of the RE's present in the chain.

For this year's QA4MRE the use of an anaphora system was needed because the analysis from last year's results revealed there was a problem in missing too many possible good excerpts. This problem could be solved using an anaphora resolution engine, since it would be possible to find antecedent references and link them, giving the QA system a better chance of finding possible excerpts. Thus the process of using the anaphora resolution engine is as follows: extract the text from each input file; pass this text to a pre-processing chain that involves: sentence splitting, tokenization, pos-

tagging and lemmatization, noun phrase chunking; after this pre-processing the resulted tagged data is passed to RARE, which outputs the co-reference chains; a post-processing tool comes and reads RARE's output, and changes the original input file, such as the references appear solver (change the referent with the antecedent); this new input file containing anaphora resolution is then passed to the QA system.

3 Results and Evaluation

For the QA4MRE 2013 task, our team submitted 5 runs, all of which were for the Romanian-Romanian language pair.

As was the case for the past QA4MRE editions, the evaluation of the results was carried out in two different manners: on a global level and on a document specific level. At the global level, the purpose of the evaluation is to provide a general measure for the quality of the system in a general setting, on any type of background knowledge. As an important note, the global evaluation does not include the evaluation of the auxiliary questions. At the document level, the purpose is to determine the value of the "c@1 measure", which is a description of how well a given text is "understood" by the QA system. The c@1 measure is also computed at the topic level. These results are then used to obtain statistical measures, such as the mean, average and standard deviation over values grouped by topic or as an overall view.

In the tables below we give the results obtained by four of our five submitted runs for Romanian, each representing a specific configuration of the system. In the case of the first two runs (C1 and C2), the system configuration included the Textual Entailment module. The difference between the runs is given by the difference in choosing the threshold for providing the "NOA" response. Our intent was to evaluate the impact of a more permissive configuration, which gives less "NOA" answers versus a more restrictive one. For the last two runs (C3 and C4), the Textual Entailment module was not used; the difference between them also comes from different NOA thresholds. The final run yielded identical results to the first one (using a similar architecture and different NOA thresholds) and is not included in the tables because of this.

3.1 Evaluation at the question answering level

Table 1 below gives the results for the four runs described above.

Table 1: Results of UAIC's Ro-Ro runs at question answering level

	C1	**C2**	**C3**	**C4**
Answered right	45	68	44	36
Answered wrong	117	202	211	149
Total answered	162	270	255	185
Unanswered right	24	1	7	15
Unanswered wrong	96	11	22	84

	C1	C2	C3	C4
Unanswered empty	2	2	0	0
Total unanswered	122	14	29	99
Overall accuracy	**0.16**	**0.24**	**0.15**	**0.13**
C@1 measure	**0.23**	**0.25**	**0.17**	**0.17**

As can be seen in Table 1, the best result of our system in terms of both overall C@1 measure and overall accuracy is obtained for the run in which the Textual Entailment module was used, together with a lower threshold for the unanswered questions (which leads to the system submitting more answers). The major drawback of the decreased threshold for submitting an answer is the fact the large majority of unanswered questions are, in fact, questions for which the system did not extract the correct answer. For example, decreasing the threshold for the C2 run meant that all the correct answers extracted by the system were actually submitted (22 answers) but that 85 more wrong answers were also submitted, greatly decreasing the general accuracy.

3.2 Evaluation at the reading test level

In Table 2, we present the median and mean for the C@1 measure for each of the 4 topics, Topic1 (Alzheimer), Topic2 (Music and Society), Topic3 (Climate Change) and Topic4 (AIDS) and their overall values for the Ro-Ro runs.

Table 2: Results of UAIC's Ro-Ro runs at reading test level

	C1	C2	C3	C4
Topic 1 median	0.21	0.22	0.14	0.14
Topic 2 median	0.14	0.17	0.17	0.17
Topic 3 median	0.19	0.22	0.25	0.26
Topic 4 median	0.25	0.28	0.07	0.08
Overall median	**0.21**	**0.22**	**0.12**	**0.13**
Topic 1 average	0.22	0.28	0.18	0.17
Topic 2 average	0.14	0.23	0.16	0.16
Topic 3 average	0.19	0.22	0.24	0.24
Topic 4 average	0.27	0.28	0.08	0.08
Overall average	**0.22**	**0.25**	**0. 17**	**0.16**

These results in term of average and median are consistent with the trend introduced in Table 1. The best overall average was obtained on the second run, which was also the highest scoring in terms of overall accuracy and C@1 measure. As can be seen above, the lowest scoring topic (in the best scoring run) is the second one (Music and society), and we have chosen to carry out our error analysis on this topic alone.

3.3 Error analysis

In extension to the analysis carried out above, we have also performed an error analysis over the reported results. The analysis was carried out exclusively over the questions in topic 2 (the lowest scoring topic in terms of C@1 measure in the second run, which yielded the best results), and a report of the most relevant error sources is given below. In interpreting the analysis results, two important factors need to be taken into account:

- *Firstly*, the analyzed run is obtained using the textual entailment enhanced system. In order to use the TE engine, we generated 5 hypotheses out of the initial question and its 5 potential answers (the potential answer was included in the hypothesis according to a set of patterns depending on the question and expected answer type).
- *Secondly*, the run for which we have chosen to carry out error analysis was obtained using a lower threshold for submitting an answer, which resulted in a large number of incorrect answers being reported.

A major source of errors for our system was the fact that we did not properly model the questions for which one of the answer variants was "none of the above". In those cases where the correct answer is note "none of the above", our system is slightly favored, because a query generated from this phrase scores very low because the keywords are rarely found in the target text. This is the case for question 3, reading 5 topic 2,

```
Ro: În ce manieră se arată influenţa lui Beethoven în sonate?

En: How is Beethoven's influence seen in the sonatas?
```

However, in those cases where the correct answer is indeed "none of the above", our system has almost no chance to extract the correct answer. In the case of question 2, reading 5, topic 2,

```
Ro: La ce dată a debutat Cramer ca dirijor?

En: When did Cramer begin his conducting career?
```

the query generated from the correct variant does not return any snippets. Because of this the score for this query is lowest, and the correct answer cannot be chosen. In order to prevent this type of error, we propose that if all the queries score less than a given threshold (which we need to find experimentally), the selected variant should be "none of the above".

One type of error which we have encountered frequently is due to the way in which we create two sets of queries in order to make use of the provided variants while searching for the correct answer. In the case of question 1, reading test 5, topic 2, the

answer our system has selected was answer one, while the correct solution was number 4.

Ro: Cât sunt de renumite sonatele târzii ale lui Cramer în ziua de azi?

En: How well known are Cramer's sonatas today?

Even though the initial query scores first, it is penalized by the secondary query, which scores third and thus misses the correct answer. The problem in such error cases usually stems from the fact that the correct answer as too few keywords, and therefore it is harder for such a query to score high enough to be picked. We propose boosting the scores of answer variants with fewer words in order to correct these types of errors.

Another type of common error is caused by the fact that, because of computational and time restrictions, we have chosen to not perform significant preprocessing on the indexed text. Because of this, especially in the case of the Romanian language, which is heavily inflected, some queries score extremely low because the words in the query are not in the same inflection as the words in the original text. This is the case for question 5, reading 5, topic 2:

Ro: Ce organizaţie l-a numit director pe Cramer în 1822?

En: Which organization made Cramer its director in 1822?

where the original text contains the keywords in a different inflection: "Academia Regală de Muzică" vs. "Academiei Regale de Muzică". If this difference in inflection would have been detected, the system would have chosen the correct answer. To this end we are considering the preprocessing of all the background knowledge.

4 Conclusions

This paper presents the updated Question Answering system developed by UAIC for the Machine Reading Evaluation task within CLEF 2012 labs. The presented systems were built starting from the main components of our QA systems (the question processing and information retrieval modules), but the multiple choice questions were addressed using a textual entailment component.

The evaluation shows a best overall median for all 4 topics of 0.22. We can observe the influence of the correctly unanswered questions in the C@1 measure when comparing the number of right answers for the best run (C2), with the run C1. Although in the C2 run, a higher number of questions were correctly answered (68 right answers) than in the C1 run (45 right answers), the C@1 measure obtained for

the C1 run (0.23) is very close to C2 run (0.25). This is explained by the difference in the number of correctly unanswered questions: 24 for C1 and only 1 for C2.

Acknowledgement. The research presented in this paper was funded by the project MUCKE (Multimedia and User Credibility Knowledge Extraction), number 2 CHIST-ERA/01.10.2012.

References

1. Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Sporleder, C., Forascu, C., Benajiba, Y., Osenova, P.: Overview of QA4MRE at CLEF 2012: Question Answering for Machine Reading Evaluation. CLEF 2012 Evaluation Labs and Workshop Working Notes Papers, 17-20 September, 2012, Rome, Italy (2012)
2. Iftene, A., Gînscă, A.L., Moruz, A., Trandabăţ, D., Husarciuc, M., Boroş, E.: Enhancing a Question Answering system with Textual Entailment for Machine Reading Evaluation. Notebook Paper for the CLEF 2012 LABs Workshop - QA4MRE, 17-20 September, Rome, Italy (2012)
3. Iftene, A., Balahur-Dobrescu, A.: Textual Entailment on Romanian. The third Workshop on Romanian Linguistic Resources and Tools for Romanian Language Processing. ISSN 1843-911X, pp. 109-118, 14-15 December. Iasi, Romania. (2007)
4. Iftene, A., Balahur, A.: Answer Validation on English and Romanian Languages. In Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. Lecture Notes in Computer Science, vol. 5706/2009, pp. 385-392. (2009)
5. Gînscă, A. L., Boroş, E., Iftene, A., Trandabăţ, D., Toader, M., Corîci, M., Perez, C. A., Cristea, D.: Sentimatrix - Multilingual Sentiment Analysis Service. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-WASSA2011). Portland, Oregon, USA, June 19-24. (2011)
6. Iftene, A., Balahur-Dobrescu, A.: Named Entity Relation Mining Using Wikipedia. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). 28-30 May, Marrakech, Morocco. (2008)
7. LUCENE: http://lucene.apache.org/java/docs/.
8. Orăsan, C., Cristea, D., Mitkov, R., Branco, A.: Anaphora Resolution Exercise – An Overview. In Proceedings of LREC-2008, Marrakech, Morocco. (2008)
9. Cristea, D., Dima, E. G.: An integrating framework for anaphora resolution. In Information Science and Technology, Romanian Academy Publishing House, Bucharest, vol. 4, no. 3-4, pp. 273-291. (2001)