Question Answering on English and Romanian Languages

Adrian Iftene[1], Diana Trandabăţ[1, 2], Ionuţ Pistol[1], Alex-Mihai Moruz[1, 2], Maria Husarciuc[1, 2], Mihai Sterpu[1], Călin Turliuc[1]

[1] UAIC: Faculty of Computer Science, "Alexandru Ioan Cuza" University, Romania
[2] Institute for Computer Science, Romanian Academy Iasi Branch
{adiftene, dtrandabat, ipistol, amoruz, mhusarciuc, mihai.sterpu, calin.turliuc}@info.uaic.ro

Abstract. This year marked UAIC[1]'s fourth consecutive participation at the QA@CLEF competition, with continually improving results. This year we participated successfully both in Ro-Ro task and in En-En task. A brief description of our system is given in this paper.

1 Introduction

In 2009, the QA@CLEF track was called ResPubliQA[2]. The structure and the aims of the task remain almost the same: given a pool of 500 independent questions in natural language, participating systems must return the answers for each question. The main difference from past editions comes from the fact that the answer must be a passage and not the exact answer. Another change was the document collection which in 2009 was the JRC-Acquis corpus. Both questions and documents are translated and aligned for a subset of the official languages (at least Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish).

This year we, the team working at the "Al. I. Cuza" University of Iasi, Romania, continued to improve our system built for competition from 2008 [1].

We indexed the corpora at both paragraph and document level, and we kept both types of returned snippets; if the search for the answer in paragraph snippets is unsuccessful, we try to identify the answer in documents snippets.

The general system architecture is described in Section 2, while Section 3 is concerned with presentation of results. Last Section presents conclusions regarding our participation in QA@CLEF 2009.

[1] "Al. I. Cuza" University
[2] http://celct.isti.cnr.it/ResPubliQA/

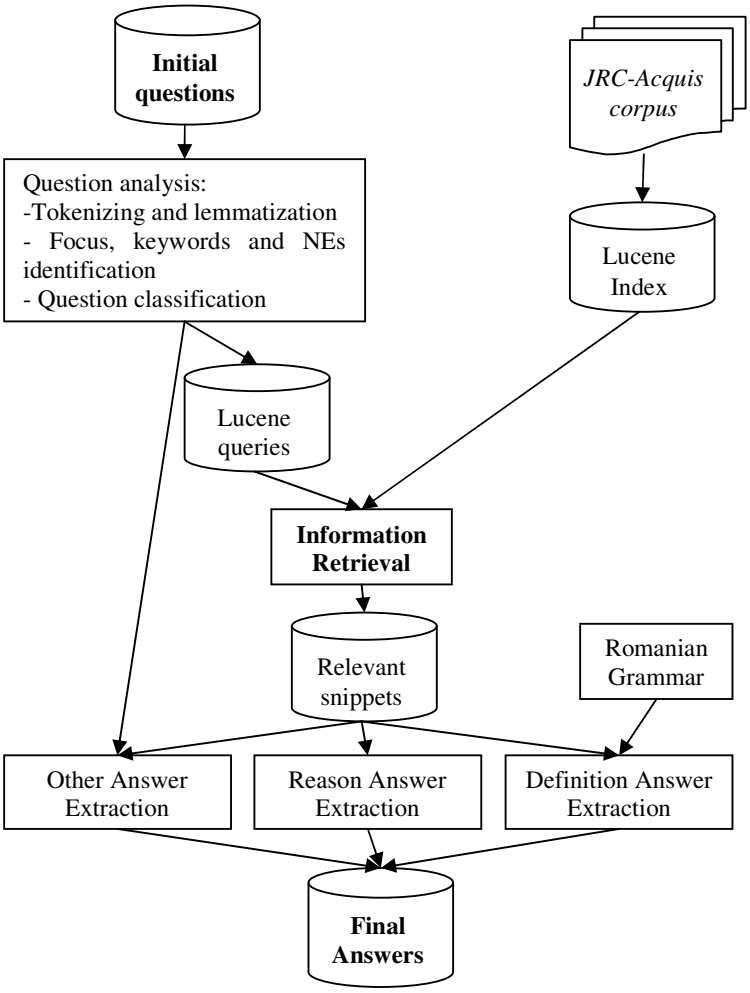

Figure 1: UAIC system used in ResPubliQA

2 Architecture of the QA System

The system architecture is similar to that of our previous systems (see Figure 1). Similarly to last year, we eliminated many pre-processing modules in order to obtain a real-time system. Also, we use a Romanian grammar in order to identify the definition type answers, and special attention was given to reason questions.

2.1 Corpus Pre-processing

The JRC-Acquis corpus, with which we had to work with for this year's competition is given in XML format, with each paragraph already marked. Therefore, since we do not do any sort of pre-processing other than paragraph separation, the pre-processing step for this corpus was not required.

2.2 Question Analysis

This step is mainly concerned with the identification of the semantic type of the answer (expected answer type). In addition, it also provides the question focus, the question type and a set of relevant keywords. The question analyzer performs the following steps:

i. NP-chunking and Named Entity extraction;
ii. Question focus identification;
iii. Answer type identification;
iv. Question type inferring
v. Keyword generation.

2.3 Index Creation and Information Retrieval

The purpose of this module is to retrieve the relevant snippets of text for every question. For this task we used the Lucene[3] indexing and search tools. Below we have given a brief description of the module:

i) Query creation

Queries are created based on the question analysis, as described in section 2.2. They are made up of the sequences of keywords we previously identified, which are modified using some of the Lucene operators, such as score boosting (the "^" operator, followed by a positive integer), fuzzy matching (the "~" operator, followed by a number greater than 0 but less than 1) and the "exclusive or" operator (symbolized by words between parentheses). As a rule of thumb, the score for the question keyword is boosted by a factor of 2 (^2), the score for the named entities in the query is boosted by a factor of 3 (^3), and, in the case of words that are not in a lemma form, we use the "exclusive or" operator between the surface and the base forms (the inflected form is boosted by a factor of 2, however).

As Romanian is a heavily inflected language, in order to avoid using all of the inflected forms of a given word, and also to avoid lemmatizing the entire corpus, we have used the fuzzy match operator, which searches for words that a similar to a given degree to the query word. After testing, we have found that the value which gives the best results is a similarity score of 0.7. For example, in the case of the question "*La ce se referă rezoluția Consiliului despre educația copiilor lucrătorilor care se mută?"* (En: "*What is the scope of the Council resolution with regards to the children of moving workers?*"), the query is:

[3] Lucene: http://lucene.apache.org/

```
(referă^2 referi) rezoluția~0.7 Consiliului^3 despre
educația~0.7 copiilor~0.7 lucrătorilor~0.7 care (mută^2
muta)
```

ii) Index creation

The index was created on the basis of the XML files in the JRC-Aquis corpus. As the texts were already marked at the paragraph level, there was no need for a pre-processing step. We have created two indexes, one at paragraph level and one at document level; the paragraph index is more precise in terms of relevant text, and is preferred for snippet extraction. If however, the answer is not found in the paragraph index, the query is applied to the document index instead.

iii) Relevant snippet extraction

Using the queries and the indexes, and the Lucene query engine, we extract a ranked list of snippets for every question.

2.4 Answer Extraction

For this year's track we built special modules in order to extract answers for DEFINITION and REASON questions.

Our algorithm for answer extraction is based on the Lucene scores. For example, let's take the following case: we have 10 answers, each with it's Lucene score. After applying all filtering criteria, some of them have the same score. It is very likely that the paragraph with de biggest Lucene score is the correct answer (although this is not guaranteed).

Of course, there are the refinement filters, which increase the Lucene score in the following cases:

- the paragraph has the focus;
- the paragraph has some of the name entities (directly proportional with the number of these name entities);
- if the question answer type is Person, or Organization, etc., we try to identify these types of name entities in the extracted paragraphs (and increase the Lucene score accordingly with the number of them);
- if the question type is Definition, then we prefer answers with definition form, identified by our grammar [2].

After applying these criteria, all paragraphs are awarded a score and the paragraph with the biggest score is chosen.

Example 1: The name entities appear in the answer list

Question: *At what age did Albert Einstein publish his famous Theory of Relativity?*
Answer 1: *Theory of Relativity, a theory that ...*
Answer 2: *Suppose Einstein would rather ...*
Answer 3: *... Albert Einstein, which, at only 22 years old , ...*

Not only does Answer 3 has 2 name entities (while Answer 2 has only one), but it also has a numeric piece of data (22), which is automatically found by our number searcher and, for this reason, it receives more points.

Example 2: The focus appears in the answer list
Question: *What is the ozone layer?*
Answer 1: *Ozone layer is by definition a layer of ...*
Answer 2: *... to destroy the ozone layer*

As we can see, *ozone*, the focus of the question, appears also in the second answer. But, the question type is Definition. So for definitions the answer should be searches starting with the focus. So, we award extra points for the position at which the focus is found. Also, note that for the first answer we also have a definition context, introduced by the verb "is", for which we also add extra points to the final score.

3 Results

Our team submitted runs this year for two language pairs, for the ResPubliQA track: English and Romanian. The best runs results are shown in Tables 1 and 2.

Table 1: Results of UAIC's best runs

	RO-RO	**EN-EN**
answered	500	447
right	236	243
wrong	264	204
unanswered	0	53
right	0	18
wrong	0	35
empty	0	0
c@1 measure	**0.47**	**0.54**

In addition this year's competition allowed and scored unanswered questions. If a question was marked as unanswered in a submitted run but an answer was provided, that correct answer was partially scored for the final run score, encouraging not providing answers with a low "confidence" score. Each provided answer was evaluated as being *right* or *wrong*, and the unanswered questions were also allowed to be *empty*. The evaluation method and the track requirements were significantly different from those of past years, so a direct comparison between our previous results and this year's scores is difficult. However, the elimination of the answer extraction step (by requesting paragraphs as answers and not exact answers) did have a major impact in the improvement of our scores, as according to our estimates, this step accounted for about 24.8 % of our previous errors.

For the two languages we sent runs we tried two different approaches, for the *En-En* considering a confidence level below which answers were marked *unanswered*.

Choosing to ignore this for the other language has had an impact on our score (0.07 better for the *En-En* run), if we would have marked all questions as answered we would have got a slightly lower score for the *En-En* run (0.522).

4 Conclusions

This paper presents the Romanian Question Answering system which took part in the QA@CLEF 2009 competition. The evaluation shows an overall accuracy of 47 % on RO-RO and 54 % on EN-EN, which are our best result from 2006 till now.

Two major improvements were carried out this year: first we continued to eliminate the most time-consuming modules from the pre-processing step. Secondly, important improvements were made regarding the information retrieval module, where Lucene queries were built in a specific way for Definition and Reason questions. Also, we use a Romanian grammar in order to extract answers for definition questions.

Other reasons for our good results obtained this year, comes from elimination of two important sources of errors. First was determined by the fact that this year the corpus was the JRC-Acquis corpus in XML format and we didn't need any pre-processing of initial corpora (last year in 10 % of cases we pre-process in a wrong way the initial corpus). The second was determined by the fact that this year we don't need to extract the exact answer from extracted paragraphs (last year in 24.8 % we select the wrong answer from the correct extracted paragraph).

Acknowledgements

This paper presents the work of the Romanian team in the frame of the PNCDI II, SIR-RESDEC project number D1.1.0.0.0.7/18.09.2007.

References

1. Iftene, A., Pistol, I., Trandabăț, D.: UAIC Participation at QA@CLEF2008. *In Proceedings of the CLEF 2008 Workshop*. 17-19 September. Aarhus, Denmark. (2008)
2. Iftene, A., Trandabăț, D. and Pistol, I.: Grammar-based Automatic Extraction of Definitions and Applications for Romanian. In Proc. of RANLP workshop "Natural Language Processing and Knowledge Representation for eLearning environments". ISBN 978-954-452-002-1, pp. 19--25, September 26, 2007, Borovets, Bulgaria (2007)