UAIC participation at QA@CLEF2008

Adrian Iftene[1], Ionuț Pistol[1], Diana Trandabăț[1, 2]

[1] UAIC: Faculty of Computer Science, “Alexandru Ioan Cuza” University, Romania
[2] Institute for Computer Science, Romanian Academy Iasi Branch
{adiftene, ipistol, dtrandabat}@info.uaic.ro

Abstract. This year marked UAIC[1]'s third consecutive participation at the QA@CLEF competition, with continually improving results. The most significant changes to our system with regards to last year is the partial transition to a real-time QA system, consequences being the simplification or elimination of principal time-consuming tasks such as linguistic pre-processing. A brief description of our system and an analysis of the errors introduced by each module are described in this paper.

Keywords: Question Answering, Romanian Grammar.

1 Introduction

Question Answering systems, especially the real-time variety, seem to come more and more frequently in the attention of researchers. The search for more advanced web searches and the emergence of the Semantic Web seem to be the main motivations for most groups of QA research of today.

We, the team working at the “Al. I. Cuza” University of Iasi, Romania, developed our first QA system as part of our participation in QA@CLEF 2006 competition (Puşcaşu et al., 2006) where we took part in the RO-EN track. The results (9.47 % accuracy) were poor, but this served as a well learned lesson in what a performing QA system should be able to do and gave us some new ideas as to how. In the 2007 competition, the CLEF organizers introduced the Romanian Wikipedia as a Romanian language corpus, thus it became possible for us to take part in the RO-RO QA track. We scored better than the first year (12 %) (Iftene at al., 2008), but the most significant improvement was the streamlining of the full QA system serving as the base of what would become this year's participation.

For this year's Romanian corpus was the same as in QA@CLEF2007, the November 2006 frozen version of the Romanian Wikipedia. This year we decided to make the task a bit harder for ourselves by trying to transform our system as close as possible to real-time QA. Due to this reason we eliminated most time-consuming pre-processing steps (POS[2] and NEs[3] identification) and we kept at minimum the number of tools involved in this part. This proved to not have a major impact on our results, as they are significantly better than last year's.

The second important improvement was regarding information retrieval part, where Lucene queries were built in a specific way for Definition questions, and the searches were done in files with the same title as the entity that must be defined. We indexed the corpora in two ways: at paragraph level and at document level, and we kept both types of returned snippets. If the search of the answer in paragraph snippets is without success, we try to identify the answer in documents snippets.

The last main improvement was done at the answer extraction part, where we tried to build very specific patterns in order to identify the final answer. For example, the MEASURE type was divided in three subtypes SURFACE, LENGTH, and OTHER_MEASURE. In this way, we improved the quality of the extraction module by specialising the patterns used. Also, in order to extract for definitions questions, we use a specialised grammar.

The general system architecture and the most important modules are described in the next section, with special focus on newly inserted components. The third section makes an analysis of our system's errors.

[1] “Al. I. Cuza” University
[2] Part-Of-Speech
[3] Name Entities

2 QA System Architecture

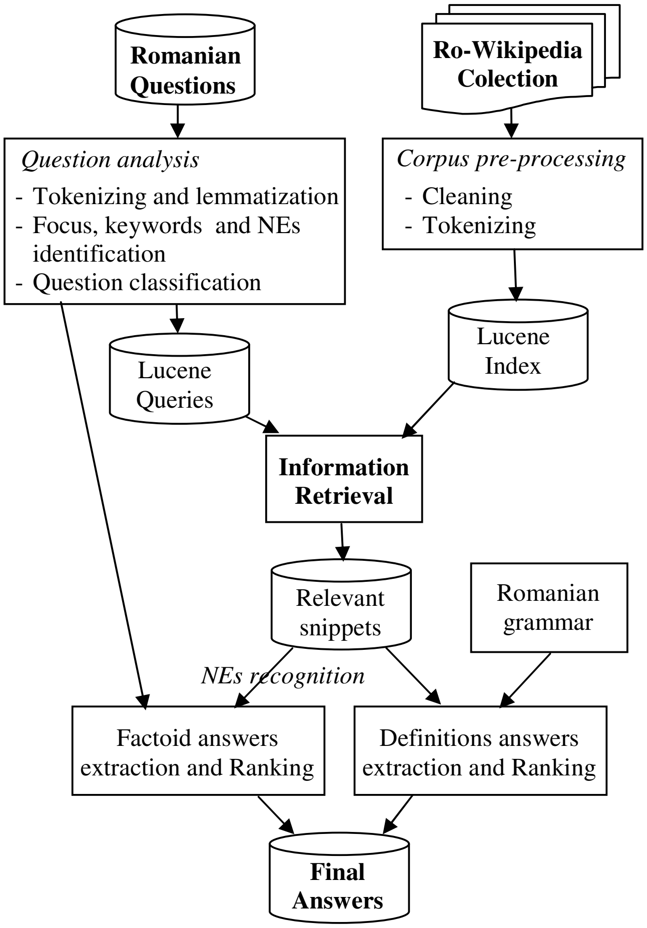

Figure 1: UAIC system used in QA@CLEF2008

The system architecture is similar to our previous systems (see Figure 1). Only few remarks: we eliminate the POS identification from pre-processing modules, we use a Romanian grammar in order to identify the definition answers, and the execution of the NEs recognition over the corpora was done only on relevant snippets in order to reduce the time necessary for system execution.

2.1 Corpus Pre-processing

The Wikipedia set of Romanian documents serving as this year's Romanian CLEF corpus comprises from 180.500 html files with a total size of 1.9 GB. The documents include topic related information, as well as forum discussions, images and user profiles. The first step prior to indexing the documents was a filtering of irrelevant information, in order to improve querying the set of documents by reducing the overall size. This was accomplished by:

- Removing documents containing images, user profiles and forum discussions. The filtering was performed automatically using pattern for the name of documents.
- Removing all of the *html* markups. The only annotated information preserved in the indexed documents is the page title and paragraph information.

These two steps reduced to corpora to 63712 documents totalling 116 MB of text. This reduction significantly reduced indexing and query search time.

Unlike our previous participations, this year we dropped the linguistic pre-processing of the corpora, as we want to make a system ready to real-time challenges. Adding linguistic information to the documents might improve the accuracy of query answers and answer extraction, but it significantly increases the time required for pre-processing and makes real-time analysis inefficient.

2.2 Question Analysis

This stage is mainly concerned with the identification of the semantic type of the entity sought by the question (expected answer type). In addition, it also provides the question focus, the question type and the set of keywords relevant for the question. This year to achieve these goals, our question analyzer performs the following steps (we specify at every step the improvements from this year):

i) NP-chunking and **Named Entity extraction** modules are the same like in previous systems.

ii) Question focus: additional rules specify to skip the first noun in some cases: when first noun is *city*, or *country*, etc. The motivation for this comes from the fact that usually at questions like "*În ce oraş s-a născut Vladimir Ilici Lenin*?" (En: *In what city was born Vladimir Ilici Lenin*?) the answer is in the fragments of text in which the word "*oraş*" (En: "city") is missing, like "*Lenin s-a născut în Simbirsk, Rusia.*" (En: *Lenin was born in Simbirsk, Rusia*).

iii) The answer type: the answer type was divided in more specific types. Thus, the PERSON name entity type was divided in PERSON, MALE and FEMALE, the MEASURE name entity type was divided in SURFACE, LENGTH, and MEASURE, LOCATION in CITY, COUNTRY, REGION, RIVER, OCEAN and LOCATION. The nature of these changes comes from fact that the tool used for NEs identification (GATE) uses the same specific sub-types, and this change will help us in the answer extraction part.

iv) Inferring the question type - the same like last year

v) Keyword generation: For keywords we consider the focus, verbs, nouns and NEs. For all these words we consider the form from question and also the word lemma.

vi) Anaphora resolution: For grouped questions we use an anaphora resolution system in the following way:

- We insert a reference to the previous question's answer;
- We add to the current list of keywords all NEs from the previous queries from the same group of questions.
- After we run the QA system, we replace the reference to the question answer with the answer itself, and we perform the information retrieval and answer extraction steps again, for the current question.

2.3. Index Creation and Information Retrieval

The purpose of this module is to retrieve the relevant snippets of text for every question. The purpose of this module is to retrieve the relevant snippets of text for every question. For those tasks, we used Lucene[4] indexing and search facilities. Below is a brief description of the module:

i) Query creation

Queries are created using the sequences of keywords, Lucene mandatory operator "+" and relevance operator "^" and "title" field for definitions questions. In this manner we obtain a regular expression for every question, which is then used in the search phase. If a word is not preceded by any operator, then that word is optional and the rank of the snippet containing that word is not boosted because of it. The words between brackets are connected by the logical operator XOR, which means that exactly one of them should be found in a snippet in order for it to be returned by the retrieval module. For example at question with id = 0001 "*Câte zile avea aprilie înainte de 700 î.Hr.*?" (En: *How many days had April before 700 î.Hr.*?) the Lucene query is

```
+(zile^2 zi) aprilie^3 700^3 î.Hr.^3
```

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

How we can see the initial form of the word receive a greater relevance like it lemma (2 for “zile” (En: days) instead of 1 for “zi” (En: day)), and the names entities receive also a greater relevance (3 in comparison with the rest).

Another case is for definitions questions where is used the “title” field. For example at question “*Cine este Ares*?” (En: *Who is Ares*?) the Lucene query is

```
(title:Ares) Ares
```

Of course in this case a greater score will receive the files with title of the document “Ares” that contain also the word “Ares”.

ii) Index creation

We have created the index of the document collection using the document tokens determined in the pre-processing phase. We have created two indexes, one at paragraph level and one at document level.

Index creation at the paragraph level

The main purpose of this type of indexing is to identify and retrieve a minimum useful amount of information related to a question. Of course the advantage is that from a reduced amount of information, we could easier identify and extract the answer from the retrieved paragraph.

Index creation at document level

An alternative indexing method was indexing at article level. The disadvantage of this method is the larger quantity of text being retrieved by a query, thus more refined answer extraction algorithms were necessary.

iii) Relevant paragraph extraction

Using the queries and the index, we extracted with Lucene a ranked list of articles / paragraphs for every question.

2.4. Answer Extraction

The retrieving process depends on the expected answer type: the answer retrieval module identifies the named entities in every snippet provided by Lucene and matches them to the answer type. When the answer type is not an entity type name, the answer retrieval syntactic patterns are based on the question focus.

For identification of the name entities we use GATE[5], which is able to identify the following types of name entities: JOB, PERSON (Family, Male, and Female), LOCATION (Region, City, Country, River, and Ocean), ORGANIZATION, and COMPANY. For the MEASURE and DATE types we start from the patterns built in competition from 2007, and complete and split them into subtypes. Thus, we split the MEASURE pattern that was a number followed by a measure unit in three patterns: LENGTH, SURFACE and OTHER_MEASURE. For LENGTH pattern we consider numbers followed by one of the following unit measures: *kilometru, metru* and *centimetru* (En: kilometre, metre, and centimetre) singular, plural and short form (km, m and cm). For SURFACE we consider the same pattern from LENGTH and we put the condition to be followed by one from surface indicators 2 or ^{2} or *pătrați* (En: *square*). The rest of unit measures were added in the rest of OTHER_MEASURE pattern (A, kHz, Hz, radian, rad, ohm, grade, kelvin, K, N, joule, J, watt, W, volt, V, min, mg, pascal, Pa, etc.). The same split operation was done for DATE type where we consider YEAR and FULL_DATE.

This year we built a special module in order to extract answers for DEFINITION questions. This module is based on a Romanian grammar (Iftene et al., 2007) built for the LT4eL project[6]. Definitions have been categorized in six types in order to reduce the search space and the complexity of grammar rules. The types of definitions observed in Romanian texts have been classified as follows:

1. “**is_def**” – Definitions containing the verb “este” (En: is):
2. “**verb_def**” – Definitions containing specific verbs, different by “este” (En: is). The verbs identified for Romanian are “indica” (En: denote), “arăta” (En: show), “preciza” (En: state), “reprezenta” (En: represent), “defini” (En: define), “specifica” (En: specify), “consta” (En: consist), “fixa” (En: name), “permite” (En: permit).

[5] http://gate.ac.uk/

[6] http://www.let.uu.nl/lt4el/

3. “**punct_def**” – Definitions which use punctuation signs like the dash “-”, brackets “()”, comma “,” etc.
4. “**layout_def**” – Definitions that can be deduced by the layout: they can be included in tables when the defined term and the definition are in separate cells or when the defining term is a heading and the definition is the next sentence.
5. “**pron_def**” – Anaphoric definitions, when the defining term is expressed in a precedent sentence and it is only referred in the definition, usually pronoun references.
6. “**other_def**” – Other definitions, which cannot be included in any of the previous categories. In this category are constructions which do not use verbs as the introducing term, but a specific construction, such as “i.e.”.

3. Results

The evaluation of the Romanian system participating in the CLEF@QA2008 competition revealed the results presented in table 2:

Table 2: Official results

Result evaluation		
Z	UNKNOWN	0
R	CORRECT	62*
U	UNSUPPORTED	4
W	WRONG	125
X	INEXACT	9*
	TOTAL	**200**

Each answer was evaluated as being UNKNOWN (unevaluated), CORRECT, UNSUPPORTED (no supporting snippet provided), WRONG or INEXACT (incomplete answer). The precision of our system was 31 %, with 19 % better like the accuracy obtained last year.

The main improvement was done at questions of type definition where we got 17 correct answers, in comparison to 0 in 2007. A more detailed analysis of our results can be provided when the official “golden” answers are provided.

4 Conclusions

This paper presents the Romanian Question Answering system which was enrolled and participated in CLEF 2008. The evaluation shows an overall accuracy of 31%, which is our best result yet (since 2006, our first participation).

Three main improvements were done this year. First we eliminated the modules most time-consuming from the pre-processing part. Secondly important improvement was regarding information retrieval part, where Lucene queries were built in a specific way for Definition questions using the title field. The last main improvement was done at answer extraction part, where we built very specific patterns in order to identify the final answer. Also, we use a Romanian grammar in order to extract answers for definition questions. As a further development, we will keep the separation in very specific sub-types for question types, because this helps very much the answer extraction module.

* In the official evaluation, the Romanian evaluator decided to penalize us because some of the answers were provided in a different order than the order of the questions, although they were correctly referenced as answers to the right question. This leads to an official result of 45 correct answers and 26 inexact answers, where all correct answers given in another order than the question order were marked as inexact.

The significant improvements shown this year combined with the major reduction in the processing time required by our system show promise regarding our goal, which is to migrate towards real-time QA.

Acknowledgements

This paper presents the work of the Romanian team in the PNCDI II, SIR-RESDEC project number D1.1.0.0.0.7/18.09.2007 and within the CEEX Rotel project number 29.

Special thanks goes to the other members of the Romanian team in the QA@CLEF competition, Dan Cristea, Alex Moruz, Maria Husarciuc, Cristina Popovici and Anna Metzak, involved in some stages of the development effort.

References

Giampiccolo, D., Forner, P., Peñas, A., Ayache, C., Cristea, D., Jijkoun, V., Osenova, P., Rocha, P., Săcăleanu, B., Sutcliffe, R. 2007. *Overview of the CLEF 2007 Multilingual Question Answering Track*. In Alessandro Nardi and Carol Peters (eds.) Working Notes for the CLEF 2007 Workshop, 19-21 September, Budapest, Hungary.

Iftene, A., Trandabăţ, D. and Pistol, I. 2007. Grammar-based Automatic Extraction of Definitions and Applications for Romanian. In Proceedings 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.

Iftene, A., Trandabăţ, D., Pistol, I., Moruz, A., Balahur-Dobrescu, A., Cotelea, D., Dornescu, I., Drăghici, I., Cristea, D. 2008. UAIC Romanian Question Answering system for QA@CLEF. In CLEF 2007. C. Peters et al. (Eds.), Lecture Notes in Computer Science, LNCS 5152, Pp. 336-343, Springer-Verlag Berlin Heidelberg 2008

Puscasu, G., Iftene, A., Pistol, I., Trandabăţ, D., Tufiş, D., Ceauşu, A., Stefănescu, D., Ion, R., Dornescu, I., Moruz, A., Cristea, D.: Cross-Lingual Romanian to English Question Answering at CLEF 2006. CLEF 2006, Revised Selected Papers, *Lecture Notes in Computer Science* vol. 4730/2007, pp. 385-394.