The DI@UE's participation in QA4MRE: from QA to multiple choice challenge

José Saias and Paulo Quaresma

Departamento de Informática, ECT
Universidade de Évora, Portugal
{jsaias,pq}@di.uevora.pt

Abstract.
This QA4MRE edition brought two challenges to the DI@UE team: the absence of Portuguese as a working language and the different nature of the task when compared with previous participation in QA@CLEF. We addressed this multiple choice answering problem by assessing answer candidates in a text surface based manner, without a deep linguistic processing. This system employs a Lucene based search engine and Wordnet to assist in synonym check and morphological normalization.

Answer analysis and the criteria for the answering decision are fundamentally based on superficial analysis of document text, enriched with semantic validation of compatibility between terms.

The solution we describe answered to 73 from 120 questions, having 18 correct answers and an accuracy of 0.15.

1 Introduction

This paper describes the participation of the Informatics Department of the University of Évora (DI@UE) team at Question Answering for Machine Reading Evaluation (QA4MRE)[1] of Cross Language Evaluation Forum (CLEF2011). In previous editions of CLEF, DI@UE focused on the Portuguese monolingual Question Answering (QA) task[2]. In QA@CLEF 2008 we used the Senso system [4] for open domain QA, featuring a portuguese stemmer, a text indexation engine and an answer validation module. In this system's latest evolution[5], candidate answers for each question are contextualized over time, space and semantic dimension. This organization enables a multiple perspective differentiation over the answer list and supports the answer appreciation process, but it is not multilingual. It incorporates tools for the Portuguese language, which was not included in the list of languages for this QA4MRE edition. Even with the introduction of tools for the English language, Senso system did not seem the most suitable for processing QA4MRE questions.

[1] http://celct.fbk.eu/QA4MRE/

[2] University of Évora previous work at CLEF: 2004 [7], 2005 [8], 2007 [3] and 2008 [4].

The task in which participated until 2008 was substantially different from that required this year. For QA@CLEF, the purpose was to automatically find the answer to a set of questions. Systems were required to detect the possible answers by themselves inside the document collections (Wikipedia and news corpus) [9].

QA4MRE main task aims to test systems' ability to understand the meaning communicated by a text [10]. The task proposes a reading comprehension exercise where each question about a document will have five choices, from which systems will identify the correct answer. Despite the complexity of the process, with the need for a justification with the elements that support the answer and the need for textual inference, these task rules give rise to specialized approaches. Using the benefit of having an answer among five possibilities, the effort can be directed to assessing answer candidates. Such may be seen as a subtask of *full QA*, in the sense that it does not need an answer extraction phase.

The next section presents the main resources employed and the system architecture. The approach used in QA4MRE is described with examples in section 3. The obtained results are presented in section 4. Finally, some conclusions and future work are pointed out in section 5.

2 System Resources and Architecture

The main system components and their interactions are presented in figure 1. The XML Layer is a component responsible for parsing the input, sorting out the questions with their multiple choice answers and maintaining a connection to their particular reading test document. When all questions are processed, this component generates the XML output and makes sure the syntax is correct and conforms to the DTD.

The Question Classifier module was thought to determine the type of the question, which is later considered in assessing each response. The Local KB has a starting knowledge base containing common sense facts about places, entities and events. Its content is important for example in Named Entity Recognition (NER) process.

The Libs Module contains collections of text documents, refered as Background Collections (BC). The English version for all three topics (*AIDS; Climate Change; Music and Society*) corresponds to approximately 2 gigabytes of text. This module also includes a Lucene[3] based text search engine, used to index all BC text and subsequent document retrieval operations. Working with English written text, Wordnet[2] is a significant external tool for assisting in synonym check, term base form normalization or to find definitions. This resource is consulted through the *Java API for WordNet Searching*[4].

[3] Apache Lucene is an open source project with advanced indexing and searching features. http://lucene.apache.org/

[4] Java API for WordNet Searching: http://lyle.smu.edu/˜ tspell/jaws/index.html

The Answer Analyzer is responsible for assess each answer choice for a question. This check is fundamentally based on superficial analysis of text, enriched with semantic validation of compatibility between terms. Besides the reading test document content, the system also examines BC documents that are retrieved from the question text and answer choice's text. Possibly relevant phrases are highlighted for further consideration when deciding between the various choices. With the information collected for each candidate answer to a question, the Answer Selector module applies a set of criteria to choose the most plausible answer. Next section explains how the system processes each QA4MRE question.

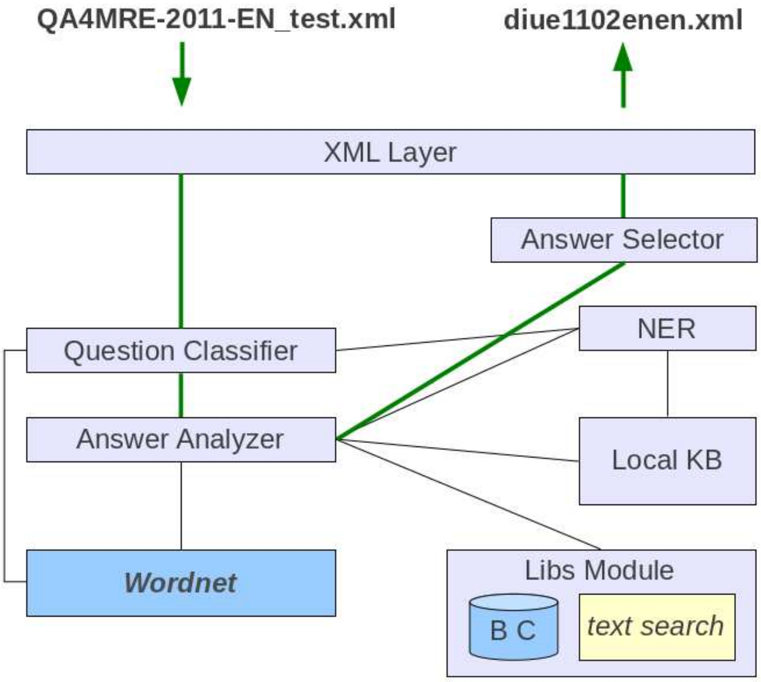

Fig. 1. System Architecture

3 Methodology

Although Machine Reading can be defined as the automatic understanding of text [6], our participation in QA4MRE is not based on a deep linguistic processing. Our approach to identify the correct answer choice to a question comprises the following steps:

- **Named Entity Recognition** - prior identification of any entity names, dates, quantities or other expressions that influence question interpretation.
- **Question Classification** - determine the category of the question in order to adopt specific procedures in the treatment of answers. The question: "*How many orchestras were mentioned in the London Times?*" is a *Quantity* subtype of *Factoid* category.
- **Document Retrieval** - search for documents in Background Collections that can support one of the possible answers to the question. It uses the Lucene tool with a search expression according to the question category and each answer candidate. Search expression avoids stop words and can be expanded using synonyms or morphological normalization.
- **Passage Retrieval** - to minimize text area where the more time consuming techniques must be applied, the retrieved documents and the reading test document are divided into text segments.
- **Answer Analysis** - all possible answers are analyzed in the text segments. For each of the multiple choices we intend to verify:

  A- Is there a textual answer pattern to this question category that is verified, in the current text segment, for this answer choice?

  B- If both question and answer key elements are present in the segment, what is the (minimal) distance between them?

  In both cases, the question key element to find in the text segments is the question focus. That is the entity or object that the question refers to, about which some information is to be determined. Question focus is identified with the category of the question. When question classification fails, the default procedure is to look for terms in the question text, filtering out stop words. The presence of those key elements on a text segment is based on term exact match, firstly, but also through semantic compatibility (synonym, hyperonym, base form).
- **Answer Selection** - considering the cases A and B detected for each answer, the system will identify the most plausible answer. The decision is based on the following ordered criteria:
  1. If there is one single answer that verified the case A, then that response is chosen.
  2. If multiple answers verify the case A:
     (a) If one of them has more occurrences of the case A, then that answer is chosen.
     (b) Otherwise, between the answers having the same number of occurrences of case A, the tie is resolved by the criteria that follow.
  3. If there is one single answer that verified the case B, then that answer is chosen.
  4. If multiple answers verify the case B and one (only) of them has the minimal distance observed in a text segment, then that answer is chosen.
  5. If multiple answers verify the case B and one (only) of them has the minimum value for the medium distance observed in the segments where it had occurences, then that answer is chosen.
  6. If none of the above applies, then the question remains unanswered.

The fifth criterion is the main difference between the two submitted runs. For the first run, it was used only to break ties resulting from the criterion 4, whereas in the last run it is applied more broadly.

4 Results

This methodology was applied to the 120 questions. Two full executions were completed and their results were submitted. For the second run, the chart in figure 2 denotes the proportion of correct and incorrect answers on the one hand, and the amount unanswered questions on the other side. It makes clear that the approach leads to a large number of wrong answers. However, if we add the right answers to the unanswered questions we get more than the number of wrong answers.

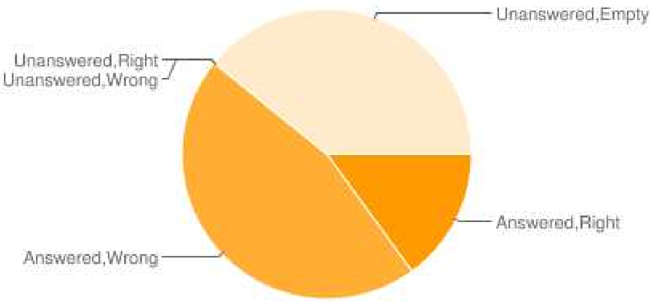

Fig. 2. Evaluation at QA level for the second run

A more specific assessment in each of the runs can be found in table 1. We can see that the system answered 74 questions in the first run. The remaining 46 questions were left unanswered without selecting candidate answer. In cases where response was submitted, only 15 were classified as correct. In the second run, one more question was left unanswered. The system was correct in 18 of the 73 responses submitted. We note that the system classification for the second run is more favorable, because while it increased the number of correct answers, it also decreased the number of erroneous results. This improvement is confirmed by the measure in the second column on the right of the table. The accuracy is calculated as the number of responses submitted and classified as correct divided by the number of questions. On run 02, accuracy is $18/120 = 0.15$, which is better than the former.

C@1 is a balanced measure rewarding systems that, for the same number of correct answers, perform better over the remaing answers [1]. By leaving some questions unanswered, a system can decrease the number of incorrect results.

	unanswered				answered			all	
Run	**#**	**Right**	**Wrong**	**Empty**	**#**	**Right**	**Wrong**	**Accuracy**	**C@1**
01	46	0	0	46	74	15	59	0.13	0.17
02	47	0	0	47	73	18	55	0.15	0.21

Table 1. QA level evaluation for each run

This measure is calculated using the formula in equation 1. The C@1 value for each run is shown in the rightmost column of the table 1. Again, the system got the best result in the second run, with C@1 = (18+47(18/120))/120 = 0.21.

The next section has some conclusions about these results and considerations about our QA4MRE participation.

$$C@1 = \frac{\#correct + \#unanswered * \frac{\#correct}{\#questions}}{\#questions} \tag{1}$$

5 Discussion

We believe that the outcome of this task is not comparable with our previous work in QA@CLEF. The aim of this work is to automatically understand the meaning of each question and its response hypotheses in order to determine the answer.

We found that answer analysis with variant A did not contributed to any answer. This may be due to problems with the question classifier, who managed to correctly assign a category to only 9 questions and some of those still had problems in detecting the question focus. Thus, the criteria 3, 4, 5 and 6 turned out to be the dominant to the process of multiple choice answering.

Looking at the questions and despite our text surface based approach can be much improved, it seems that there is a limit beyond which only a deeper semantic analysis may reach the answers.

This QA4MRE edition brought us two challenges: the absence of Portuguese as a working language and the different nature of the task and its objectives. Having a lack of time to implement a more robust solution, we consider that the results are satisfactory. Moreover, we believe this experiment constitutes a basis for a future semantically more advanced system, enabled to English, that can be tested in an upcoming participation.

References

1. Anselmo Peñas and Alvaro Rodrigo. *A Simple Measure to Assess Non-response.* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, pages 1415–1424, (2011), ISBN: 978-1-932432-87-9.

2. George A. Miller. *Wordnet: A lexical database for English.* Communications of the ACM, (1995)
3. José Saias and Paulo Quaresma. The senso question answering approach to portuguese qa@clef-2007. Technical report, CLEF 2007 Working Notes, Cross-Language Evaluation Forum Workshop, Budapest, Hungary, (2007). ISBN: 2-912335-32-9.
4. José Saias and Paulo Quaresma. The senso question answering system at qa@clef 2008. Technical report, Universidade de Évora, Multiple Language Question Answering @ Cross-Language Evaluation Forum, (2008). ISBN: 2-912335-43-4.
5. José Saias. *Contextualização e Activação Semântica na Selecção de Resultados em Sistemas de Pergunta-Resposta.* Phd Thesis, (2010), hdl.handle.net/10174/2505
6. Lucy Vanderwende. *Answering and Questioning for Machine Reading.* American Association for Artificial Intelligence, (2007)
7. Paulo Quaresma, Luis Quintano, Irene Rodrigues, José Saias and Pedro Salgueiro. *The University of Évora approach to QA@CLEF-2004.* CLEF 2004 Working Notes, (2004)
8. Paulo Quaresma and Irene Rodrigues. *A Logic Programming Based Approach To QA@CLEF05 Track.* CLEF 2005 Working Notes, (2005)
9. QA@CLEF2008. *Guidelines for the participants in QA@CLEF 2008.* http://clef-qa.fbk.eu/2008/download/QA@CLEF08_Guidelines-for-Participants_new.pdf
10. QA4MRE@CLEF2011. *Track Guidelines.* http://celct.fbk.eu/QA4MRE/