DI@UE in CLEF2012: question answering approach to the multiple choice QA4MRE challenge

José Saias and Paulo Quaresma

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

Abstract.
In the 2012 edition of CLEF, the DI@UE team has signed up for Question Answering for Machine Reading Evaluation (QA4MRE) main task. For each question, our system tries to guess which of the five hypotheses is the more plausible response, taking into account the reading test content and the documents from the background collection on the question topic.
For each question, the system applies Named Entity Recognition, Question Classification, Document and Passage Retrieval. The criteria used in the first run is to choose the answer with the smallest distance between question and answer key elements. The system applies a specific treatment for certain factual questions, with the categories *Quantity*, *When*, *Where*, *What*, and *Who*, whose responses are usually short and likely to be detected in the text. For the second run, the system tries to solve each question according to its category. Textual patterns used for answer validation and Web answer projection are defined according to the question category.
The system answered to all 160 questions, having found 50 right candidate answers.

1 Introduction

In the 2012 edition of Cross Language Evaluation Forum (CLEF), the Informatics Department of the University of Évora (DI@UE) team has signed up for Question Answering for Machine Reading Evaluation (QA4MRE)[1] main task. This was the second year that we participated in a CLEF Lab using English as working language for the system. In previous work for QA@CLEF [1, 2], we focused in Portuguese, but this language was not available in QA4MRE.
The objective is to solve questions that are given in the form of multiple choice, each having five options, and only one correct answer [3]. For the main task of

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

this year there are 160 questions, 40 more than last year [4]. Background collections have received more texts beyond those that already existed for the previous three topics, and a new topic appeared, with Alzheimer related documents.
We kept the approach used in 2011 [5], making a few adjustments to certain types of questions. Instead of objectively seek an answer to each question, as we would do in a regular QA process, we focus on assessing answer candidates. The textual justification for each answer, requested in 2011 [6], is no longer required in the system results, which allows selection techniques independent from the reference corpus.
The system architecture and the employed resources are described in the next section. The third section presents the methodology used for question processing, including some examples. Section 4 lists the results of the system, and the last section is devoted to the analysis of results and considerations on the work done.

2 System Resources and Architecture

The system maintains the architecture defined for the previous year participation. In Figure 1 we can see the different system modules. The XML Layer receives the input, does the parsing and organization of the questions with their multiple choice answers, while maintaining the connection to the topic to which they relate and 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 conforming to the DTD.
The Question Classifier module makes a linguistic analysis to the question text to determine its category. The system uses the parser CandC with the Boxer[2] semantic tool, that can produce Discourse Representation Structures (DRS) [7]. Question type is later considered for assessing each response in a more specific and targeted way. Currently, we focus on factual response categories, such as quantities, dates, names and short descriptions.
The Libs Module contains the Background Collections (BC), which include the English version of the four 2012 topics (*AIDS; Climate Change; Music and Society; Alzheimer*) documents. This corresponds to almost 3 gigabytes of text. For text retrieval we keep using the Lucene[3] search engine. Wordnet [8] is the resource used for synonym and hyperonym check, definitions and morphological normalization, consulted through the *Java API for WordNet Searching*[4].
The Local KB has a starting knowledge base containing common sense facts about places, entities and events. It can assist in Named Entity Recognition (NER) process or compatibility validation of terms for very specific cases not covered by Wordnet.

[2] http://svn.ask.it.usyd.edu.au/trac/candc/wiki/boxer
[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 includes a linguistic analysis of the text of the response and a textual search process. The search can be simple or defined according to the category of question, through textual patterns for answer projection over the BC or over the Web. With the information collected for each candidate answer to a question, the Answer Selector module applies a criteria to choose the most plausible answer. These criteria may be more general, if the question was not classified in any specific category, or more directed and concrete, as we shall see in the next section.

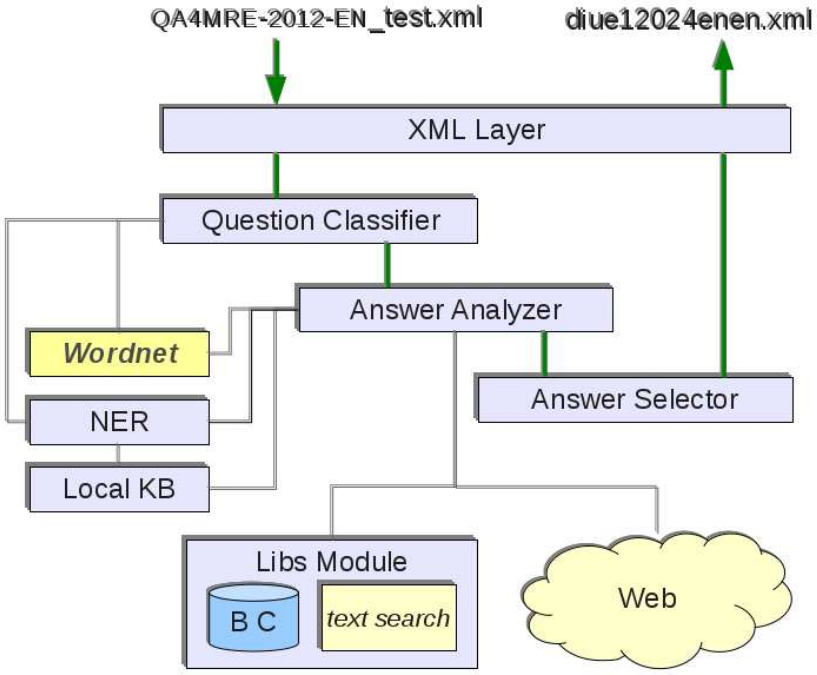

Fig. 1. System Architecture in 2012

3 Methodology

For each question, our system tries to guess which of the five hypotheses is more plausible response, taking into account the reading test content and the documents from the background collection on the question topic (Aids, Climate change, Music and Society, and Alzheimer). In this edition, we submitted two runs for QA4MRE evaluation. The first run applies a generic strategy of

surface text analysis on the reading test and documents retrieved from background collections, with no other external resources. For each question, the process begins with Named Entity Recognition, for prior identification of any entity names, dates, quantities or other expressions that influence question interpretation. Then, processing continues with question classification, document retrieval and passage retrieval. The system performs a search for documents in the BC that can support one of the possible answers to the question.

Each candidate answer has a set of retrieved passages, which are text segments. For each of the multiple choices we intend to verify if both question and answer key elements are present in the text segments, and what is the distance between them. The criteria used in the first run is to choose the answer with the smallest distance between question and answer key elements. The question key element to find in the text segments is the question focus, the entity or object that the question refers to. In both cases, the stop words are filtered from key elements. Because the textual justification was no longer needed, we decided to answer all questions using the broader heuristic whenever a specialized strategy was not applicable.

Questions with different nature are processed differently. The system applies a specific treatment for certain factual questions, whose class and focus are identified. Such specialized strategy is applied to factoid questions with the categories *Quantity*, *When*, *Where*, *What*, and *Who*, whose responses are usually short and likely to be detected in the text. Question classification requires a question text analysis, in particular the identification of the interrogative term and the question focus. As an example, the question '*How many degrees did Burney receive from Oxford?*' has *Quantity* category. Its focus is *Burney*, and what we need about it is the number of degrees received from Oxford. In another example, the interrogative used in question '*Where was Burney working when he first conceived the idea of writing a music history?*' determines its classification in the type *Where*, guiding further processing to a location.

The use of textual patterns is very common in question answering mechanisms, as in [9], [10] or [11]. In previous QA work, we used patterns in Senso system [2], but tuned to the Portuguese language. Answer scenarios explored in that system patterns are not directly applicable to English sentence structure. Therefore, since 2011 we have been adjusting the textual patterns to identify possible factual responses.

For the second run, the system tries to solve each question according to its category, adopting a specialized strategy of resolution. The textual patterns used for answer validation are defined according to the category, representing common cases for that type of question. The system checks the presence of question and answer key elements on a text segment based on term exact match as done for the first run, firstly, but also through semantic compatibility (synonym, hyperonym, base form). This is a semantic query expansion of the search terms.

Without the need to associate with each answer a supporting text, it becomes possible to use also background collections independent heuristics, such as Web answer projection to validate the options for the question. This technique is only

used when the question is classified as one of the factoid categories supported by the system. An example where the Web answer projection helped to identify the correct choice is '*What is the population of Brazil?*'. The correct answer appears on the reading test, but away from the question focus terms. Projecting the answer choices on the web, the answer *180 millions* emerged as the right one.
The decision to choose the answer to a question is based on the following:

1. The criteria used in the first run is to choose the answer with the smallest distance between question and answer key elements.
2. The second run aims for a more informed choice process. If the system finds some question category answer pattern in a retrieved document, or on the Web, for a single answer option, then that option is the chosen. Even if other options have results for the general text surface search, the answer patterns under the question category have priority.
3. If there is a tie, in any case, the choice falls on the option having more support text segments, which are the retrieved text passages where answer patterns or key search terms were found.
4. If the surface term search, for the general case, or the answer pattern search does not return any results, for any of the five hypotheses, then the question remains unanswered.

In the following section we discriminate the results obtained in each run submitted to QA4MRE.

4 Results

In the first approach, the system answered to 156 questions using the surface based technique. Figure 2 illustrates the proportion between the amount unanswered questions, and also the correct and the mistaken results. The system gave 45 correct answers and the remaining 111 were wrong. The questions referred

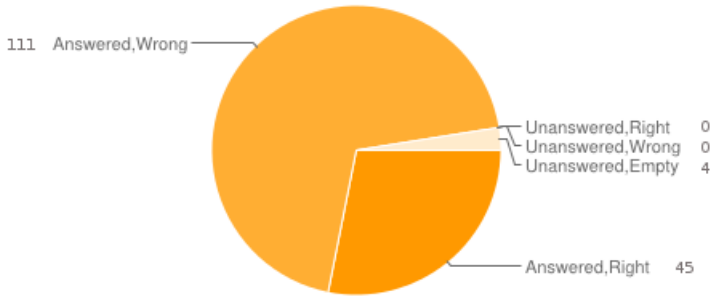

Fig. 2. Evaluation at QA level for the first run

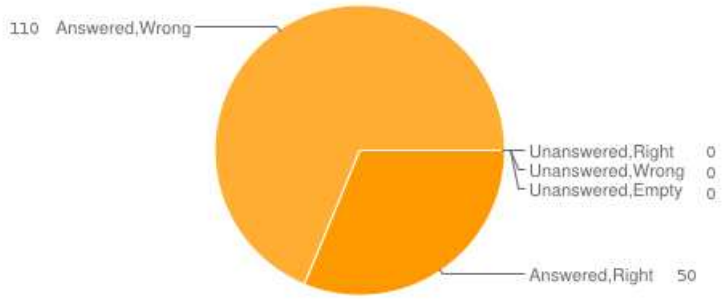

Fig. 3. Evaluation at QA level for the second run

to in the previous section have been answered, but the choice was not right[5] in any of them.

In the first run 4 questions were unanswered due to an unexpected parse error. That problem was fixed for the second run. In this last run, which is the most complete, the system changed the response in 20 of the 156 questions previously answered with the surface text search approach.

The system kept the answer to both *Burney* questions mentioned in section 3, keeping the erroneous choice. But for '*Which pupil of Dr John Blow taught Charles Burney?*', a *What* class question, the system changed from a wrong answer to the right answer (*Edmund Baker*), in the second run.

In other examples, such as '*How many countries have acted effectively against AIDS?*' and '*Where is the epicenter of the AIDS pandemic?*', were answered correctly in both runs.

The chart in Figure 3 summarizes the evaluation results for our system. The number of hits increased to 50. At the same time, 4 more questions have been processed, leaving no question unanswered, and the number of wrong answers was even reduced by one.

Of the four unanswered questions in the first run, three were answered incorrectly. Only the first of these, '*Name two styles which have contributed to pop music:*', got the right answer on the second run.

Table 1 shows a more detailed assessment with the breakdown of values for each run. The system first attempts resulted in a 0.28 accuracy and 0.29 C@1 values. C@1 is a balanced measure rewarding systems that, for the same number of correct answers, decrease the number of incorrect results by leaving some questions unanswered [12].

The accuracy rose to 0.31 in the second run, and the overall C@1 measure was also 0.31.

Table 2 shows a comparison between the result of this year's better run and the best result achieved by our system in 2011, where there were 120 questions

[5] According to the solutions disclosed in July 2012: QA4MRE-2012-EN_GS.xml.

for processing. The last section has some thoughts on these results and on our participation in this Lab.

	unanswered	answered			all		
Run	#	#	**Right**	**Wrong**	#	**Accuracy**	**C@1**
01	4	156	45	111	160	0.28	0.29
02	0	160	50	110	160	0.31	0.31

Table 1. Detailed evaluation

	unanswered	answered			all		
Year/Run	#	#	**Right**	**Wrong**	#	**Accuracy**	**C@1**
2011 best	47	73	18	55	120	0.15	0.21
2012 best	0	160	50	110	160	0.31	0.31

Table 2. Comparison with the results of previous participation

5 Discussion

In 2011, our system answered to 73 of 120 questions, finding correct 18 answers. In this edition, our system answered correctly to 50 questions out of 160. This represents a substantial improvement in accuracy, from 0.15 to 0.31.
Compared with the previous year, the question classifier has improved, being more effective in assigning the category of factoid questions. We believe that this update in the question classifier was the key to improving outcomes, especially for allowing the application of specific procedures for each category of question. This year we also improved the text analysis performed on the question and answer hypotheses, using the CandC and Boxer tools.
Looking at the charts in figures 2 and 3, the wrong answers slice is significantly higher. Answer all questions may not have been a good decision. In future we can introduce a confidence factor or more appropriate criteria to decide between responding and non-responding. With such a procedure, the system can improve the C@1 measure.
Errors in question classification not always determine a wrong answer. Question '*What is the external debt of all African countries?*' asks for a monetary value and should have been classified as *Quantity*. The system classified it as a *What* question, whose resolution process is not optimized for numeric values. Yet, both the surface search and the answer projection succeed, and the option chosen by the

system was the correct in both runs.
Despite the improved accuracy results, we consider that the number of wrong answers is very high. This may reflect a need to use more semantic based techniques, and perhaps to apply an intensive linguistic analysis to BC documents.
This second participation in QA4MRE helped us to adjust our system to English, focusing primarily on factual answer questions and specific categories, but with an alternative methodology for the general case.

References

1. 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.
2. 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.
3. QA4MRE@CLEF2012. *Track Guidelines.* http://celct.fbk.eu/QA4MRE/
4. Anselmo Peñas, Eduard H. Hovy, Pamela Forner, Álvaro Rodrigo, Richard F. E. Sutcliffe, C. Forascu, and C. Sporleder *Overview of QA4MRE at CLEF 2011: Question Answering for Machine Reading Evaluation.* http://clef2011.org/resources/proceedings/Overview_QA4MRE_Clef2011.pdf
5. José Saias and Paulo Quaresma. The di@ue's participation in qa4mre: from qa to multiple choice challenge. In Vivien Petras, Pamela Forner, and Paul D. Clough, editors, *CLEF 2011 Labs and Workshop: Notebook Papers*, Amsterdam, The Netherlands, 2011. ISBN: 978-88-904810-1-7.
6. QA4MRE@CLEF2011. *Track Guidelines.* http://celct.fbk.eu/QA4MRE/
7. Hans Kamp and Uwe Reyle. *From Discourse to Logic.* Kluwer, Dordrecht, (1993)
8. George A. Miller. *Wordnet: A lexical database for English.* Communications of the ACM, (1995)
9. Martin M. Soubbotin, Sergei M. Soubbotin *Use of Patterns for Detection of Likely Answer Strings: A Systematic Approach.* Text REtrieval Conference, (2002)
10. Sneiders, E. *Automated email answering by text pattern matching.* IceTAL, Lecture Notes in Computer Science, vol. 6233, pp. 381–392, Springer, (2010)
11. Sung, C.L., Lee, C.W., Yen, H.C., Hsu, W.L. *An alignment-based surface pattern for a question answering system.* IRI. pp. 172–177. IEEE Systems, Man, and Cybernetics Society (2008)
12. 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.