Enhancing a Question Answering system with Textual Entailment for Machine Reading Evaluation

Adrian Iftene[1], Alexandru-Lucian Gînscă[1], Alex Moruz[1, 2], Diana Trandabăț[1, 2], Maria Moruz[3], Emanuela Boroș[1]

[1] UAIC: Faculty of Computer Science, "Alexandru Ioan Cuza" University, Romania
[2] Institute of Computer Science, Romanian Academy Iasi Branch
[3] Center of Biblical-Philological Studies *Monumenta linguae Dacoromanorum*, "Alexandru Ioan Cuza" University, Romania
{adiftene, lucian.ginsca, amoruz, dtrandabat, mhusarciuc, emanuela.boros}@info.uaic.ro

Abstract. This paper describes UAIC[1]'s Question Answering for Machine Reading Evaluation systems participating in the QA4MRE 2012 evaluation task. We submitted two types of runs, first type of runs based on our system from 2011 edition of QA4MRE, and second type of runs based on Textual Entailment system. For second types of runs, we construct the *Text* and the *Hypothesis*, asked by Textual Entailment system from initial test data (the <documents> tag was used to build the *Text* and the <question> and <answer> tags were used to build the *Hypothesis*). The results offered by organizer showed that second type of runs were better than first type of runs for English.

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

1 Introduction

As in the 2011 campaign, the Question Answering for Machine Reading Evaluation (QA4MRE[2]) task in 2012 intends to cross-evaluate the ability of systems to read and understand texts. The task focuses on reading a document and identifying the correct answer from a set of five multiple choice answers, using inferences and previously acquired background knowledge. The test data and background knowledge are related to four topics: *AIDS*, *Climate Change*, *Music and Society* (the same topics adopted last year [1]) and a new one i.e. *Alzheimer* [2]. An important note is that, for all involved languages (English, Spanish, German, Italian and Romanian), the test data was the same (parallel translations) and the background knowledge was available to all participants.

For UAIC's participation in the QA4MRE task in 2012, we used as base the system built for the 2011 QA4MRE edition [3], which was, at its turn, an updated version of our previous systems from the 2009 and 2010 QA@CLEF editions [4], [5].

[1] University "Al. I. Cuza" of Iasi, Romania
[2] QA4MRE: http://celct.fbk.eu/QA4MRE/index.php

The base system was further improved by adapting a Textual Entailment component for the Question Answering module, similar to the approach in [6].

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 2012, UAIC submitted two types of runs for Romanian and English. For the first type of runs, we use the system from the previous edition of QA4MRE 2011 [3], consisting in modules specialized for *test data processing, background knowledge indexing*, *snippet extraction* and *identification of the correct answer*. For the second type of runs, we use the Romanian and the English textual entailment systems [7, 8], similar to the approach detailed in [9]. The English system is similar to the Romanian system, with the difference that a part of the modules presented in subsections 2.1 were only partially used.

2.1 The base architecture

In 2012, the Romanian background knowledge consisted of a collection of 184,263 documents in text format (28,826 correspond to the *AIDS* topic, 57,160 to *Climate Change* topic, 88,687 to *Music and Society* topic and 9,590 to *Alzheimer* topic). The test data consists in an XML file with 16 test documents (4 documents for each of the four topics), 10 questions for each document (160 questions in total) and 5 possible answers for each question (800 possible answers in total).

The base architecture is similar to the system used for the 2011 edition of the QA4MRE competition, presented in [3]. Thus, after indexing the background collection using Lucene[3] libraries, 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 [1] and the web services available both for Romanian and English from the Sentimatrix[4] project [10] to eliminate stop words, perform lemmatization and identify the Named Entity in the question. Then, a Lucene query is build. For instance, in the case of the question with q_id = "8":

Ro: *Care dintre următoarele nu este o cauză a vulnerabilității femeilor căsătorite față de infecțiile cu HIV*?
En: *Which of the following is not a cause of HIV infection for married women*?

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

the execution of the above steps has the following results:

- in the first step, the following stop words are eliminated: *care, dintre, următoarele, o, a, de, cu* (En: *which, of, following, a, for*);
- in the next step, lemmas for the words *cauză, vulnerabilității, femeilor, infecțiile* (En: *cause, vulnerability, women, infections*) are identified;
- in the third step, *HIV* is identified as a Named Entity;
- in the last step, the Lucene query is build: *"nu (cauză^2 cauza) (vulnerabilității^2 vulnerabilitate) (femeilor^2 femeie) (căsătorite^2 căsătorit) (infecțiile^2 infecție) HIV^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 [11], 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:

Ro: *În ce stat american oamenii de știință universitari au calculat că pentru combaterea SIDA în Africa fiecare american trebuie să plătească un cost de 5 dolari anual?*
En: *In what American state did university scientists calculate the cost to each American of spending 5 dollars annually to combat AIDS in Africa?*,

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

As presented in [3], the index of background knowledge is queried, and all retrieved documents are placed in separate indexes. The results of this step are 160 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*. Finally, 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.

2.2 Enhancing the base architecture with Textual Entailment

The architecture of the components that used the Textual Entailment (TE) system is presented in Figure 1, being based on the system used in AVE exercises in 2007 and in 2008 [9] and being similar to the architecture of one of the best systems from the QA4MRE 2011 edition [13].

The steps executed by our system are as follows:

- We build a pattern with variables for every question according to the question type;

- Using a pattern and all possible answers, we build a set with 5 hypotheses for each of the questions: H_1, H_2, H_3, H_4, H_5;
- 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: (T, H_1), (T, H_2), (T, H_3), (T, H_4), (T, H_5).

Finally, we consider the candidate from the hypothesis for which we obtain the greatest global fitness to be the correct answer for a current question.

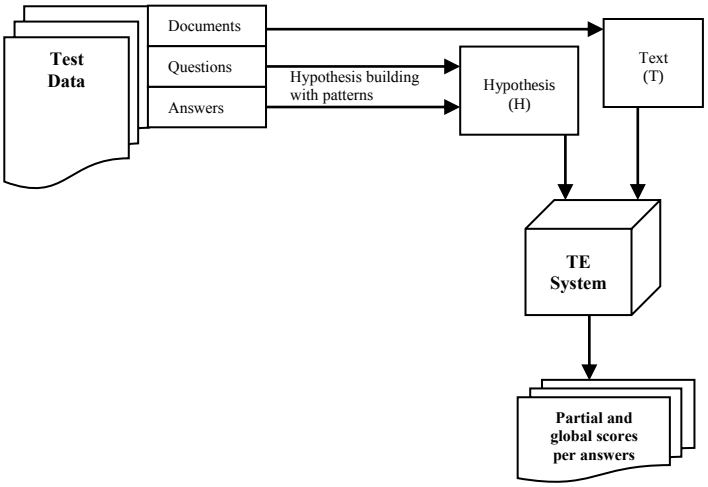

Figure 1: The second architecture based on Textual Entailment (TE) system

Pattern building

In order to use the TE system for ranking the possible answers in the QA4MRE task, all these questions are first transformed according to the algorithm presented in [14].

For example, for the following question we have:

Question: *What is the goal of the ABC strategy?*

Our program generates the following pattern:

Pattern: *ANSWER is the goal of the ABC strategy.*

where *ANSWER* is the variable in this case. We generate specific patterns according to the following answer types: Measure (*How many, How much*), Person (*Who, Name*), Location (*In what*), Date (*On what date, When*) and Other. Following the building of the pattern, we proceed to constructing the corresponding hypotheses.

Hypothesis building

Using the pattern building mechanism above and the answers provided within the QA4MRE input XML data, we built the corresponding hypotheses. For example, for the question above, we built, according to the answers from the English test data, the following hypotheses:

H_1: *promoting women's social and economic rights is the goal of the ABC strategy.*

H_2: *combating women's inequalities is the goal of the ABC strategy.*

H_3: *promoting abstinence, faithfulness, and use of condoms is the goal of the ABC strategy.*

H_4: *ignoring poverty, social inequality and traditional ways is the goal of the ABC strategy.*

H_5: *promoting the prevention and treatment programs is the goal of the ABC strategy.*

For each of these hypotheses, we consider that the corresponding text from the "document" tag as having the role of text T.

Answers classification

We consider the pairs built above as input for our Textual Entailment system. After running the TE system, the global fitness values for the exemplified pairs are the following:

GlobalFitness(H_1, T) = 2.1854732

GlobalFitness(H_2, T) = 1.3577608

GlobalFitness(H_3, T) = 1.92097

GlobalFitness(H_4, T) = 2.2404695

GlobalFitness(H_5, T) = 2.2766914

Since in the considered case the highest value is obtained for the answer 5 "*promoting the prevention and treatment programs",* we consider it as the most probable answer. The NOA answers were considered the pairs for which we have the maximum value for GlobalFitness very close to 0.

3 Results and Evaluation

For the QA4MRE 2012 task, our team submitted 10 runs, out of which 5 were for the Romanian-Romanian language pair and 5 for the English-English pair.

The evaluation of the results is done from two different perspectives in a similar manner as in the 2011 QA4MRE edition. The first one is equivalent to a traditional evaluation in which all the answers are gathered in a single set which is then compared to a gold standard, not taking into account the document associated with a particular answer. On the other hand, the reading perspective offers insight on how well the system "understands" a particular document. At first, the C@1 measures of each test comprising of 10 questions per document are taken into consideration. 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 following 4 tables, we detail the result obtained by each of the 5 different configurations for Romanian and other 5 configurations for English. In each case, the first two configurations (C1 and C2) refer to the first architecture design. The difference between C1 and C2 represents 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. The last three configurations represent runs in which the architecture involving Textual Entailment system was used. The difference between these three configurations resides, as in the case of the first two, in the different choice of threshold for the "NOA" answers. We tested a permissive, a moderate and a restrictive threshold.

3.1 Evaluation at the question answering level

In Table 1, we present the results for the 5 runs on Romanian and in Table 2, the same results are provided for the 5 runs on English.

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

	C1	**C2**	**C3**	**C4**	**C5**
Answered right	34	38	34	33	21
Answered wrong	114	111	113	104	67
Total answered	148	149	147	137	88
Unanswered right	3	3	4	0	12
Unanswered wrong	9	8	9	23	60
Total unanswered	12	11	13	23	72
Overall accuracy	**0.21**	**0.24**	**0.21**	**0.21**	**0.13**
C@1 measure	**0.23**	**0.25**	**0.23**	**0.24**	**0.19**

As can be seen in Table 1, the best result of our system in terms of C@1 measure is obtained for the run in which the first type of architecture was used together with a slightly more permissive threshold for the unanswered questions. Contrary to this, for English, two out of the three query reformulation runs outperform the best result of the first two configurations. This shift can be explained by the increased effectiveness of the patterns applied for query rewriting when working on the English language.

Table 2: Results of UAIC's En-En runs at question answering level

	C1	**C2**	**C3**	**C4**	**C5**
Answered right	34	23	34	37	25
Answered wrong	96	65	78	104	62
Total answered	130	88	112	141	87
Unanswered right	7	16	6	3	15
Unanswered wrong	23	54	42	16	58
Total unanswered	30	72	48	19	73
Overall accuracy	**0.21**	**0.14**	**0.21**	**0.23**	**0.16**
C@1 measure	**0.25**	**0.21**	**0.28**	**0.26**	**0.23**

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 for Romanian, with the best from the English runs. Although in the Ro-Ro run, a higher number of

questions were correctly answered (38 right answers) than in the En-En run (34 right answers), the C@1 measure obtained for the English run (0.28) is higher than the one given by the best Romanian run (0.25). This is explained by the difference in the number of correctly unanswered questions.

A common denominator between the results for Romanian and those for English is that a balanced threshold provided the best results. This is best observed when comparing the last three configurations both for English and for Romanian. For example, in Table 1, the C4 configuration in which there were 24 unanswered questions outperformed the C3 (13 unanswered questions) and C5 (72 unanswered questions). The same pattern is found in Table 2, for the En-En runs.

3.2 Evaluation at the reading test level

In Table 3, we present the median and mean for each of the 4 topics, Topic1 (AIDS), Topic2 (Climate Change), Topic3 (Music and society) and Topic4 (Alzheimer) and their overall values for the Ro-Ro runs. In Table 4, the same results are provided for the En-En runs.

Table 3: Results of UAIC's Ro-Ro runs at reading test level

	C1	**C2**	**C3**	**C4**	**C5**
Topic 1 median	0.16	0.26	0.26	0.24	0.18
Topic 2 median	0.31	0.31	0.31	0.20	0.21
Topic 3 median	0.20	0.20	0.15	0.18	0.07
Topic 4 average	0.29	0.29	0.21	0.28	0.21
Overall median	**0.24**	**0.29**	**0.26**	**0.23**	**0.16**
Topic 1 average	0.18	0.28	0.26	0.25	0.16
Topic 2 average	0.27	0.27	0.27	0.25	0.26
Topic 3 average	0.20	0.20	0.18	0.18	0.07
Topic 4 average	0.26	0.26	0.19	0.26	0.20
Overall average	**0.23**	**0.25**	**0.22**	**0.23**	**0.17**

Table 4: Results of UAIC's En-En runs at reading test level

	C1	**C2**	**C3**	**C4**	**C5**
Topic 1 median	0.25	0.30	0.31	0.30	0.33
Topic 2 median	0.21	0.23	0.25	0.21	0.22
Topic 3 median	0.20	0.14	0.28	0.23	0.17
Topic 4 average	0.20	0.08	0.22	0.18	0.16
Overall median	**0.21**	**0.16**	**0.29**	**0.26**	**0.22**
Topic 1 average	0.28	0.29	0.34	0.34	0.32
Topic 2 average	0.22	0.22	0.24	0.24	0.20
Topic 3 average	0.23	0.17	0.25	0.24	0.18
Topic 4 average	0.22	0.08	0.22	0.19	0.16
Overall average	**0.23**	**0.19**	**0.26**	**0.25**	**0.21**

These results in term of average and median are consistent with the trend introduced in Table 1 and Table 2. The best overall mean was obtained for the third configuration on English and the second one, on Romanian.

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 topic was arbitrarily chosen), 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 submitted runs are grouped according to the basic philosophy regarding query generation. In the first case (the first three submitted runs), queries were generated on the basis of the question alone, and then the potential answer was searched in the top scoring results. In the second case (the last two submitted runs), we generated 5 queries for each question, one for each potential answer (the potential answer was included in the query). The textual entailment system was also used in the second case.
- *Secondly*, the various runs for each case were obtained by tweaking the threshold at which the system decided to provide an answer.

One of the first types of error we have encountered is the fact that the second type of runs has different queries than the first type. This is the case of the first question in the second topic, reading 5:

Ro: *Care dintre următorii este un biocombustibil*?
En: *Which of the following is a biofuel*?

The first three runs provide the correct answers (using the method described above), while the last two run have wrong answers. This is due to the fact that the query `(următorii^2 următor) biocombustibil etanol` provides lower scoring snippets than the query `(următorii^2 următor) biocombustibil carbon` (the correct answer is "ethanol", but the provided answer is "carbon"). Upon examining the snippets returned by the five queries extracted for this particular question, we have discovered that the correct answer scored fourth overall, which practically excludes the correct answer from consideration. This type of error was also encountered for questions 4, 5 and 6, reading 5 in the 2nd topic.

A more subtle type of error is the one which generated an incorrect answer for the question 2, reading 5, topic 2. In this case, regardless of the manner of creating the query, the chance of obtaining the correct answer is low because of the nature of the base text. The fault comes from the answer extraction module, which is unable to solve coreference, and therefore cannot extract the correct answer

Ro: *... combustibilul lichid este atât de valoros. Până în prezent, este câştigătorul evident atunci când avem nevoie de energie pentru transport - în special transport aerian şi transport maritim greu pe distanţe mari - deoarece ne permite să*

înghesuim o grămadă de energie într-un spaţiu de stocare relativ mic şi să realimentăm cu uşurinţă...

Some errors are caused by the query generation module, such as the case of question 3, reading test 5, topic 2. In this instance, none of the five submitted runs provided the correct answer, mainly because most of the query words are not found in the vicinity of the correct answer. The query generated by the question analysis module is:

Ro: *(poate^2 putea) (mărită^2 mări) (cantitatea^2 cantitate) (culturi^2 cultură) (cultivate^2 cultivat) simultan bucată pământ*

to which the system then adds the query:

Ro: *(folosind^2 folosi) (culturi^2 cultură) (anuale^2 anual) (succesive^2 succesiv)*

for the expected answer. The text span which contains the correct answer is:

Ro: *A doua premiză greşită din scenariile cele mai nefavorabile este aceea că de pe aceeaşi suprafaţă de teren nu se pot obţine mai multe recolte. Amestecurile perene pentru biocombustibili celulozici ar putea fi, de fapt, cultivate alături de culturile anuale sau, pe acelaşi teren, între recoltare şi însămânţare...*

and we can see that some of the keywords of the query are not found within it. The only way in which the system could solve this is by using synonyms for the keywords (in this case, *culturi* in the question and *recolte* in the answer). The fact that most of the keywords in the query are not found in the supporting text is also highlighted by the fact that the system did not provide an answer for this question in run 5, because of the low score of all the retrieved snippets. The same type of error was observed in the case of question 9, reading 5, topic 2 and question 6, reading 6, topic 2. An extreme case of this error can be seen in question 1, reading 6, topic 2, where none of the runs gave any answer, although the extracted answer was correct, because of the los score of the supporting snippets.

In some cases, errors arise from the addition of the second query in the case of the first three runs. This can be seen for the answer generated for question 7, reading 5, topic 2, where the initial query provides high scoring snippets which contain the correct solution:

Ro: *O parte din problemă provine din apetitul deosebit al porumbului pentru stimulente, cum ar fi îngrăşămintele.*

but these snippets are then penalized because of the second query (in this particular case, the secondary query *(absoarbe^2 absorbi) (cantităţi^2 cantitate) (reduse^2 reduce) (gaze^2 gaz) efect seră* introduces a far score for a

different snippet, because of the high number of keywords compared to the correct solution query, *nevoie (cantităţi^2 cantitate) (mari^2 mare) fertilizatori^2*). This type of issue could be corrected to some extent by the use of synonymy, which would increase the score of the correct snippet. This issue can also be seen in the case of question 8, reading 5, topic 2 and questions 2 and 4, reading 6, topic 2.

A type of error that stems from the lack of sufficient background knowledge can be found for question 10, reading 5, topic 2:

Ro: *Care este biocombustibilul a cărui producţie reduce cel mai mult emisiile de gaze cu efect de seră?*
En: *Which is the biofuel which reduces greenhouse gas emissions most?)*

The correct answer for this question, *etanol celulozic* (En: *cellulose ethanol*) cannot be found as such in the text, although it is referred in another form:

Ro: *o versiune celulozică de etanol*
En: *a cellulosic version of ethanol*

as can be seen in the snippet containing the correct answer:

Ro: *Există o mulţime de moduri diferite de a face biocombustibili celulozici, inclusiv o versiune celulozică de etanol, şi ei reduc emisiile cu un procent enorm, între 82% şi 87%.*

This type of problem can only be corrected by the appropriate 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 0.29 for both the Romanian and English monolingual tasks. 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 for Romanian, with the best from the English runs. Although in the Ro-Ro run, a higher number of questions were correctly answered (38 right answers) than in the En-En run (34 right answers), the C@1 measure obtained for the English run (0.28) is higher than the one given by the best Romanian run (0.25). This is explained by the difference in the number of correctly unanswered questions.

Acknowledgement. The research presented in this paper was funded by the Sector Operational Program for Human Resources Development through the project "Development of the innovation capacity and increasing of the research impact through post-doctoral programs" POSDRU/89/1.5/S/49944.

References

1. Peñas, A., Rodrigo, A.: A Simple Measure to Assess Non-response. In Proceedings of 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies (ACL-HLT 2011), Portland, Oregon, USA, June 19-24. (2011)
2. 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)
3. Iftene, A., Gînscă, A. L., Moruz, A., Trandabăţ, D., Husarciuc, M.: Question Answering for Machine Reading Evaluation on Romanian and English Languages. Notebook Paper for the CLEF 2011 LABs Workshop, 19-22 September, Amsterdam, Netherlands. (2011)
4. Iftene, A., Trandabăţ, D., Moruz, A., Pistol, I., Husarciuc, M., Cristea, D.: Question Answering on English and Romanian Languages. In C. Peters et al. (Eds.): CLEF 2009, LNCS 6241, Part I (Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments). Pp. 229-236. Springer, Heidelberg. (2010)
5. Iftene, A., Trandabăţ, D., Moruz, A., Husarciuc, M.: Question Answering on Romanian, English and French Languages. Notebook Paper for the CLEF 2010 LABs Workshop, 22-23 September, Padua, Italy. (2010)
6. Iftene, A., Balahur-Dobrescu, A.: Improving a QA System for Romanian Using Textual Entailment. In Proceedings of RANLP workshop "A Common Natural Language Processing Paradigm For Balkan Languages". Pp. 7-14, September 26, 2007, Borovets, Bulgaria. (2007)
7. Iftene, A.: Textual Entailment (Ph.D. Thesis) Technical Report. "Al. I. Cuza" University. ISSN 1224-9327. 169 pages. October, 2009. Iasi, Romania. (2009)
8. 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)
9. 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)
10. 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)
11. 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)
12. LUCENE: http://lucene.apache.org/java/docs/.
13. Pakray, P., Bhaskar, P., Banerjee, S., Chandra Pal, B., Bandyopadhyay, S., Gelbukh, A.: A Hybrid Question Answering System based on Information Retrieval and Answer

Validation. Notebook Paper for the CLEF 2011 LABs Workshop, 19-22 September, Amsterdam, Netherlands. (2011)
14. Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini B., Szpektor, I.: The Second PASCAL Recognising Textual Entailment Challenge. In Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment. Venice. Italy. (2006)