Evaluating Answer Validation in multi-stream Question Answering

Álvaro Rodrigo   Anselmo Peñas   Felisa Verdejo
Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
Madrid, Spain
{alvarory, anselmo, felisa}@lsi.uned.es

Abstract

We follow the opinion that Question Answering (QA) performance can be improved by combining different systems. Thus, we planned an evaluation oriented to promote the specialization and further collaboration between QA systems. This multi-stream QA requires to develop the modules able to select the proper stream according to the question and the candidate answers provided. We describe here the evaluation framework we have developed with special focus on the evaluation measures and the study of their behavior in a comparative evaluation.

Keywords: *Evaluation, Question Answering, Answer Validation.*

1 Introduction

Traditional Question Answering (QA) systems typically employ a pipe-line approach, in which the traditional steps are: question analysis, document retrieval, passage selection and answer extraction [6, 9, 13]. However, this kind of architecture has a dependency among modules that is highly sensitive to error propagation. For instance, a QA system using document retrieval and answer extraction modules both of which performing with a precision of 80%, would have an upper bound precision of 64% due to the dependency between modules and error propagation. Besides, it is no clear that improving a single component also improves the overall performance of the system [14]. Therefore, it is evident that it is necessary to go beyond the pipeline processing and to promote the development of other architectures.

1.1 Promoting Collaboration

A multi-stream QA system is composed by several QA systems which receive questions and search for answers (see Figure 1). Besides, there is a subsystem which receive all the answers from the single QA systems and select one as the final answer of the multi-stream [7].

One of the possibilities for obtaining advances in the area would be to encourage the specialization of systems and the collaboration among them. This idea arises from the behavior of current systems. For example, [8] reports a QA evaluation where 81% of the questions were correctly answered by at least one system, but the best performing system only answered correctly 52.5%. In other words, a perfect selection of the output of all participant systems would have answered correctly 81% of the questions.

Moreover, the best system is not the one with the best performance for each kind of questions. Thus, it seems promising to look for the subsystems with the goal of achieving this perfect selection in a multi-stream QA architecture.

1.2 Evaluation Proposal

One of the main challenges that arises around this idea of collaboration among systems is to develop good criteria for the selection of the final answer among the candidates returned by the alternative systems [7]. Most of works focused on improving QA accuracy by combining different approaches, have used a voting scheme based on redundancy as the selection criteria [1, 2], including sometimes information about performance history [7]. On the other hand, there have also been methods that take the decision based on the answer confidence score reported by each system [3].

However, answer validation and selection can take advantage of more sophisticated approaches,

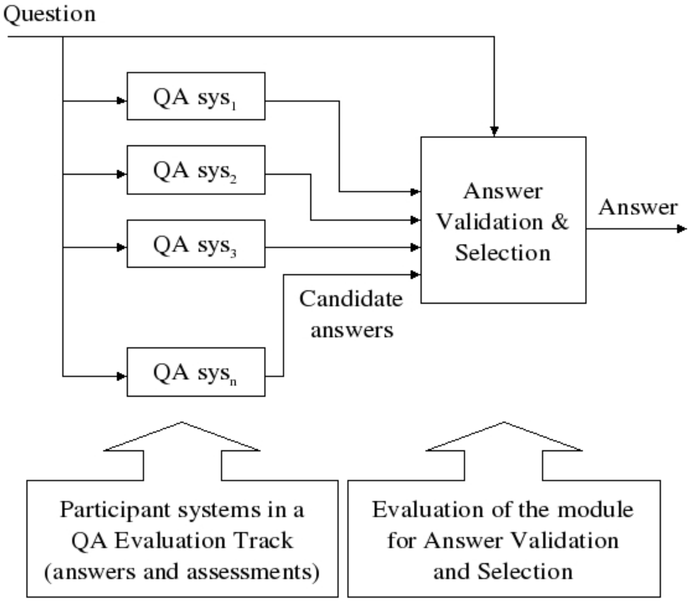

Question
QA sys$_1$
QA sys$_2$
QA sys$_3$
QA sys$_n$
Candidate answers
Answer Validation & Selection
Answer
Participant systems in a QA Evaluation Track (answers and assessments)
Evaluation of the module for Answer Validation and Selection

Figure 1. Multi-stream QA architecture and proposed evaluation

and this kind of work should be promoted. For example, a deeper analysis such as textual entailment approaches has been successfully used as a way to rank and to select answers in traditional QA [5].

In the task of Answer Validation (AV), a system receives a question, an answer to the question and a text that supports the correction of the answer. Then, the system must decide whether the answer to the question is or not correct according the given supporting text [11].

We propose to use AV systems based on sophisticated analysis as the modules inside the QA multi-stream architecture that would receive several candidate answers (from different single QA systems) and would select one as the final one given by the multi-stream.

The proposal we make here is a methodology to evaluate AV systems for selecting answers in multi-stream QA. In the paper we focused on the measures for evaluating AV systems in this scenario and we study the proposed methodology in a comparative evaluation.

1.3 Structure

In Section 2 we describe the evaluation measures required for testing AV systems and for comparing them with QA systems. In Section 3 we show a case study of the proposed methodology over the results in AVE 2007 that still not used the new measure. Finally, some conclusions are presented in Section 4.

2 Evaluation proposed

With the aim of evaluating AV systems in a multi-stream QA framework, we used different measures taking into account that AV systems can be evaluated in several ways. In our proposal we are interested in evaluating two aspects of the systems. On one hand we are interested in their ability detecting correct answers, which was the first objective for which the task of AV was proposed.

On the other hand, we are interested in evaluating AV systems selecting answers from different streams, as well as in estimating the performance that could be obtained by taking advantage of the ability of AV systems detecting questions with not correct answers. Since AV systems has not been evaluated for this purpose yet, we proposed here a new measure.

2.1 Evaluating the validation of answers

In the first group of measures the objective is to evaluate the ability of AV systems validating answers from a pool of available ones. These measures are useful for evaluating the performance of AV systems used for ranking or filtering answers returned by QA systems.

In [10] it was argued why the evaluation in AV with unbalanced collections is based on the detection of correct answers. *Precision* (1), *recall*

(2) and *F-measure* (3) (harmonic mean) over answers that must be detected as correct are used instead of using an overall accuracy as the evaluation measure. In other words, the goal is to quantify systems ability to detect whether there is enough evidence to accept an answer. Results can be compared among systems but always taking the baseline of:

- a system that accepts all answers detecting all the correct answers but dropping in precision in the same proportion of the incorrect answers in the collection.

On one hand, *precision* measure tells how good a system is when it predicts an answer as correct. It acknowledges systems that validate only correct answers.

On the other hand, *recall* measure evaluates the capability of systems for detecting all the correct answers. It acknowledges the systems ability to detect a high amount of correct answers without paying attention to the precision in this detection.

The *F-measure* permits to evaluate both aspects of a system giving the same importance to each of the measures. However, this is an intrinsic evaluation that is not enough for obtaining some evidence about the QA performance gain of using AV systems into multi-stream QA architectures.

2.2 Evaluating the selection of answers

Since the first group of measures is not able to measure the ability of AV systems selecting answers from multiple streams, the second group of measures was created with this purpose. The objective of these measures is to compare the performance of single QA systems with multi-stream QA systems that use AV for the selection of answers.

In order to perform this evaluation, AV systems receive to each question a set of answers returned by different QA systems. Then, AV systems are requested to select one answer per question when more than one has been validated. Thus, for each question there are two possible situations:

- There is only an answer selected.
- All the answers has been rejected.

Thus, it is possible to measure the selection of answers and the detection of questions without correct answers.

2.2.1 Evaluating the correct selection

Since AV systems are requested to select one or none of the answers to a question, the resulting behavior would be comparable to a QA system: for each question there is no more than one answer selected.

The first measure of this group is *qa_accuracy* (4): the proportion of questions for which a correct answer has been selected. This measure is directly comparable to the traditional accuracy used for evaluating QA systems. Therefore, we can compare multi-stream QA systems that used AV modules with single QA systems taking as reference the QA systems performance over the questions involved.

This measure has an upper bound given by the proportion of questions that have at least one correct answer. This upper bound corresponds to a perfect selection of the correct answers given by all the QA systems that take part in the multi-stream. The normalization of *qa_accuracy* with this upper bound is given by *%_best_combination* (5), where the percentage of the perfect selection is calculated.

Besides the upper bound, results of *qa_accuracy* can be compared with a random system: a system that selects randomly one answer per question. Thus, this baseline can be seen as the average proportion of correct answers per question group. We call it *random_qa_accuracy* (6).

2.2.2 Evaluating the correct rejection

qa_accuracy only acknowledges the ability of a system for selecting correct answers and not the ability of detecting that all the answers to a question are incorrect.

The justification of why to acknowledge the recognition of questions without correct answers arises from the fact that a possible gain in performance could be obtained in these questions if they are properly detected. In this situation, the AV system could ask to the QA systems for another answer to the question, opening the possibility of obtaining a correct answer to this question. In order to acknowledge this behavior, the concept of rejected questions must be introduced.

A rejected question is a question in which the AV system has not selected any answer. That means that the AV system considers that all the answers to the question are incorrect. Then, we proposed the use of *qa_rej_accuracy* (7), which acknowledges systems capable of detecting correctly rejected questions. This measure evaluates the performance of an AV system detecting the questions that do not have any correct answer.

$$precision = \frac{|VALIDATED\ correctly|}{|VALIDATED|} \tag{1}$$

$$recall = \frac{|VALIDATED\ correctly|}{|CORRECT|} \tag{2}$$

$$F = \frac{2 * recall * precision}{recall + precision} \tag{3}$$

$$qa\_accuracy = \frac{|answers\ SELECTED\ correctly|}{|questions|} \tag{4}$$

$$\%\_best\_combination = \frac{|answers\ SELECTED\ correctly|}{|questions\ with\ correct\ answers|} * 100 \tag{5}$$

$$random\_qa\_accuracy = \frac{1}{|questions|} \sum_{q \epsilon questions} \frac{|correct\ answers\ of\ (q)|}{answers\ of(q)} \tag{6}$$

$$qa\_rej\_accuracy = \frac{|questions\ REJECTED\ correctly|}{|questions|} \tag{7}$$

$$estimated\_qa\_performance = qa\_accuracy + qa\_rej\_accuracy * qa\_accuracy \tag{8}$$

2.2.3 Evaluating the potential performance

It is interesting to study the additional gain in performance that could be obtained by detecting also the questions without correct answers.

estimated_qa_performance (8) considers that the questions accounted by *qa_rej_accuracy* could be answered with the accuracy given by *qa_accuracy*. Then, this measure rewards the precision of AV systems detecting incorrect answers in the proportion they select correct ones.

3 Case of study

In order to check the viability of the proposed methodology, we applied it to the data and results of participant systems in the Answer Validation Exercise[1] (AVE) 2007 [10] at CLEF 2007.

The performing of the proposed evaluation inside the QA at CLEF allow us to have a large amount of data from real QA systems for the testing collections. Besides, the availability of data in several languages, apart from English, provided the possibility of evaluating systems in different languages such as Spanish, English, Portuguese or German.

Participant systems at QA at CLEF receive questions and they output an answer and a snippet that supports the correctness of the answer. These answers have been evaluated by human assessors (further information can be found in the QA at CLEF guidelines[2]).

With the am of applying our methodology, it was necessary to create a set of collections. These collections had to be developed in a way that allow to apply the measures of section 2 which the aim of evaluating the features of AV systems we were interested in.

Then, after applying our methodology to real systems, a look to the results helped us in obtaining conclusions about the proposed measures.

The development of the collections is described in [12]. In summary, participants at AVE 2007 received a set of questions with a group of candidate answers.

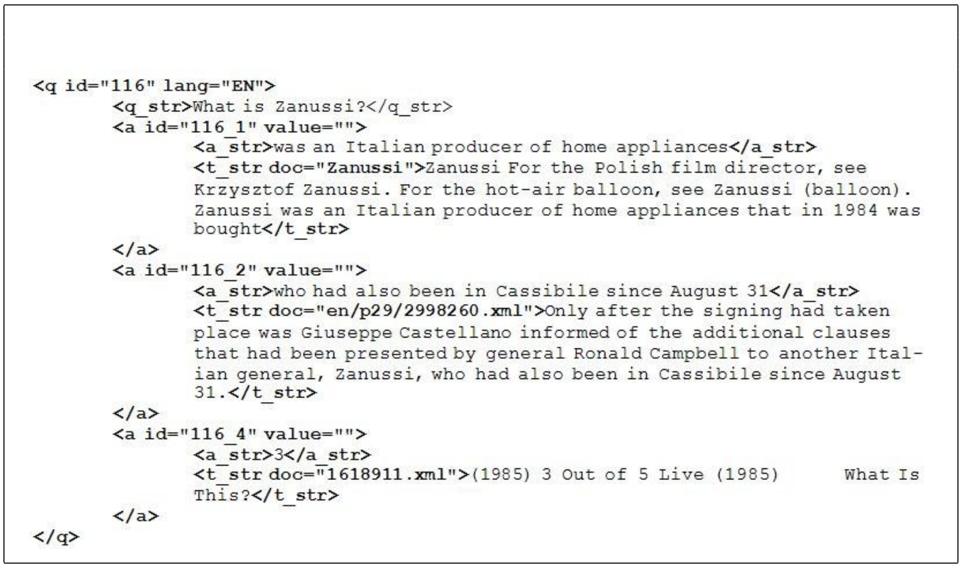
The format of the collections we obtained is similar to the one shown in Figure 2. In Table 1 it is shown the number of questions and the number of answers obtained using the output of participant systems in QA at CLEF 2007 [4] for building test collections.

3.1 Analysis of the Measures according to Results

Tables 2 and 3 (taken from [12]) show the values of *precision*, *recall* and *F-measure* over AVE 2007 participant systems in English and Spanish respectively, as well as the baseline proposed in section 2.1 for this group of measures. Systems are ranked by *F-measure* and results can not been compared among different languages due to the

[1] http://nlp.uned.es/clef-qa/ave/

[2] http://nlp.uned.es/clef-qa/QA@CLEF08_Guidelines-for-Participants.pdf



```
<q id="116" lang="EN">
        <q_str>What is Zanussi?</q_str>
        <a id="116_1" value="">
                <a_str>was an Italian producer of home appliances</a_str>
                <t_str doc="Zanussi">Zanussi For the Polish film director, see
                Krzysztof Zanussi. For the hot-air balloon, see Zanussi (balloon).
                Zanussi was an Italian producer of home appliances that in 1984 was
                bought</t_str>
        </a>
        <a id="116_2" value="">
                <a_str>who had also been in Cassibile since August 31</a_str>
                <t_str doc="en/p29/2998260.xml">Only after the signing had taken
                place was Giuseppe Castellano informed of the additional clauses
                that had been presented by general Ronald Campbell to another Ital-
                ian general, Zanussi, who had also been in Cassibile since August
                31.</t_str>
        </a>
        <a id="116_4" value="">
                <a_str>3</a_str>
                <t_str doc="1618911.xml">(1985) 3 Out of 5 Live (1985)      What Is
                This?</t_str>
        </a>
</q>
```

Figure 2. Excerpt of an English test collection.

	German	**English**	**Spanish**	**French**	**Italian**	**Dutch**	**Portuguese**	**Romanian**
Questions	113	67	170	122	103	78	149	100
Answers(final)	282	202	564	187	103	202	367	127
VALIDATED	67	21	127	85	16	31	148	45
REJECTED	197	174	424	86	84	165	198	58
UNKNOWN	18	7	13	16	3	6	21	24

Table 1. Number of questions and answers in the AVE 2007 test collections

Table 2. Precision, Recall and F-measure over correct answers for English.

System	P	R	F
DFKI_2	0.44	0.71	0.55
DFKI_1	0.37	0.62	0.46
UA_1	0.25	0.81	0.39
Text-Mess_1	0.25	0.62	0.36
Iasi	0.21	0.81	0.34
UNED	0.22	0.71	0.34
Text-Mess_2	0.25	0.52	0.34
UA_2	0.18	0.81	0.29
100% VALIDATED	0.11	1	0.19

Table 3. Precision, Recall and F-measure over correct answers for Spanish.

System	P	R	F
INAOE_1	0.38	0.86	0.53
INAOE_2	0.41	0.72	0.52
UNED	0.33	0.82	0.47
UJA_1	0.24	0.85	0.37
100% VALIDATED	0.23	1	0.37
UJA_2	0.4	0.13	0.19

different number of VALIDATED answers in each language, but they can be compared with the two baselines provided (a system that validates all the answers and a system that validates the half of the answers).

Tables 4 and 5 show the values of *qa_accuracy*, *%_of_perfect_selection* (as can be found in [12]), *qa_rej_accuracy* and *estimated_qa_performance* (that we study and compare here) over AVE participant systems and QA systems participants at QA@CLEF 2007. Systems are ranked by *estimated_qa_performance*.

Since the concept of rejected questions works only for AV systems, the values of *qa_accuracy* and *estimated_qa_performance* are equal for QA systems. The values of the baselines proposed in section 2.2 for this group of measures are also given.

The rankings using *F-measure* and the ones using *estimated_qa_performance* are different. Since they evaluate systems in a different way. The *F-measure* and the related ones of section 2.1 consider all the answers in the collections. In other words, they evaluate each answer of each stream. However, *estimated_qa_performance* and the related measures of section 2.2 consider groups of answers.

Nevertheless it is possible to see certain correlation between the *recall* measure and *qa_accuracy*. This is because traditional QA accuracy measures only accounts the correct answers, acknowledging the *recall*. A system that always selects an answer tends to increase *recall* and also *qa_accuracy*.

On the other hand, we can see that the rankings according to *qa_accuracy* and *estimated_qa_performance* are slightly different. For example in English case (see table 4), the system UA_1 performs better than DFKI_1 (0.18 against 0.16) according to *qa_accuracy*, which means that UA_1 is better selecting correct answers. However, DFKI_1 is better detecting answers without correct answers (0.43 of *qa_rej_accuracy* against 0.28 of UA_1). That is to say, UA_1 is betting for selecting more answers even incorrectly (see also that its *precision* is lower than the one of DFKI_1 according to Table 2). According to *estimated_qa_performance* this behavior is detected and DFKI_1 turns to have better punctuation than UA_1 (0.24 against 0.23). In other words, considering the possible gain in performance that might be obtained by detecting that all answers to a question are incorrect and asking for new ones, the system DFKI_1 is better. Therefore, the system DFKI_1 may help to obtain better results in QA than the system UA_1 and this is pointed out by *estimated_qa_performance*.

3.2 Analysis of the results

According to *estimated_qa_performance* there are some systems able to outperform the best QA system. This is the case of INAOE_1 in Spanish (table 5) and many systems in the case of English (table 4). This means that the use of current AV systems in multi-stream QA could lead to outperform current QA systems.

However, the result of INAOE_1 over the best QA system shows the importance of detecting questions without correct answers: if we do not consider *estimated_qa_performance* and we only consider *qa_accuracy*, the AV systems cannot beat the best QA system. In other words, the gain in performance that the AV system could provide depends not only in the correct selection of answers, but also in the correct detection of questions without correct answers.

4 Conclusions

In this paper we have presented a methodology for evaluating AV systems that select answers in multi-stream QA. For this evaluation, we have defined two different set of measures.

The first group of measures (*precision*, *recall* and *F-measure*) evaluates the validation of all answers coming from all streams. Their goal is to measure the ability of AV systems for detecting

Table 4. Comparing AV systems performance with QA systems in English.

System	System Type	qa_accuracy (% of perfect selection)	qa_rej_accuracy	estimated_ qa_performance
Perfect selection	AV	0.3 (100%)	0.70	0.51
DFKI_2	AV	0.21 (70%)	0.45	0.3
Iasi	AV	0.21 (70%)	0.31	0.27
UA_2	AV	0.19 (65%)	0.10	0.21
DFKI_1	AV	0.16 (55%)	0.43	0.24
UA_1	AV	0.18 (60%)	0.28	0.23
UNED	AV	0.16 (55%)	0.31	0.22
Text-Mess_1	AV	0.15 (50%)	0.34	0.2
UI	QA	0.18 (60%)	0	0.18
Text-Mess_2	AV	0.12 (40%)	0.37	0.16
DFKI_QA_1	QA	0.13 (45%)	0	0.13
Random	AV	0.1 (35%)	0	0.1
DFKI_QA_2	QA	0.04 (15%)	0	0.04
100% rejected	AV	0 (0%)	0.70	0

Table 5. Comparing AV systems performance with QA systems in Spanish.

System	System Type	QA accuracy (% of perfect selection)	qa_rej_accuracy	estimated_ qa_performance
Perfect selection	AV	0.59 (100%)	0.41	0.83
INAOE_1	AV	0.45 (75%)	0.20	0.54
Priberam	QA	0.49 (83%)	0	0.49
UNED	AV	0.42 (70%)	0.18	0.49
INAOE_2	AV	0.36 (61%)	0.25	0.45
UJA_1	AV	0.41 (68%)	0.01	0.41
INAOE_QA	QA	0.38 (63%)	0	0.38
Random	AV	0.25 (41%)	0	0.25
MIRA	QA	0.15 (26%)	0	0.15
UPV	QA	0.13 (22%)	0	0.13
UJA_2	AV	0.08 (14%)	0.36	0.11
TALP	QA	0.07 (12%)	0	0.07
100% rejected	AV	0 (0%)	0.41	0

correct answers and only them. Therefore, we propose them for evaluating AV systems used for validating, filtering or ranking answers.

The second group of measures evaluates the performance in the selection of a single answer from the pool given by several streams. We have shown that it is not only important to acknowledge the correct selection of answers, but also the correct detection of questions without correct answers. These two aspects are taken into account by the measure *estimated_qa_performance*.

We hope that the use of *estimated_qa_performance* will promote the development of systems that not only detect correct answers, but also the incorrect ones, improving QA performance.

Finally, according to the results obtained we have seen that it is possible to outperform state of the art QA systems by using multi-streams with current AV systems. For example, in Table 4 there are 8 systems (DFKI_2, Iasi, UA_2, DFKI_1, UA_1, UNED and Text-Mess_1) which outperforms the better QA system (UI) according to the measure *estimated_qa_performance*.

Acknowledgments

This work has been partially supported by the European project Trebe CLEF (ICT-1-4-1 215231), the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267), the Education Council of the Regional Government of Madrid and the European Social Fund.

References

[1] E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng. Data intensive question answering. In *Proc. of TREC*, 2001.

[2] J. D. Burger, L. Ferro, W. Greiff, J. Henderson, M. Light, S. Mardis, and A. Morgan. Mitre's Qanda at TREC-11. In *Proc. of TREC-11*, 2003.

[3] J. Chu-Carroll, K. Czuba, J. Prager, and A. Ittycheriah. In question answering, two heads are better than one. In *Proc. of HLT-NAACL*, 2003.

[4] D. Giampiccolo, P. Forner, J. Herrera, A. Peñas, C. Ayache, D. Cristea, V. Jijkoun, P. Osenova, P. Rocha, B. Sacaleanu, and R. Sutcliffe. Overview of the CLEF 2007 multilingual question answering track. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers, volume 5152 of Lecture Notes in Computer Science*, pages 200–236, 2008.

[5] S. Harabagiu and A. Hickl. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, 2006*, pages 905–912, 2006.

[6] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C. Lin. Question answering in webclopedia. In *Proceedings of the Ninth Text REtrieval Conference, pages 655–664*, 2001.

[7] V. Jijkoun and M. de Rijke. Answer selection in a multistream open domain question answering system. In *Proc. Of European Conference on Information Retrieval*, 2004.

[8] B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, V. Jijkoun, P. Osenova, A. Peñas, P. Rocha, B. Sacaleanu, and R. Sutcliffe. Overview of the CLEF 2006 multilingual question answering track. In *Evaluation of Multilingual and Multimodal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers, volume 4730 of Lecture Notes in Computer Science*, pages 223–256, 2007.

[9] D. Moldovan, S. Harabagiu, M. R. Pasca, M., G. R. Girju, R., and V. Rus. The structure and performance of an open-domain question answering system. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 563–570, 2000.

[10] A. Peñas, Á. Rodrigo, V. Sama, and F. Verdejo. Overview of the Answer Validation Exercise 2006. In *Evaluation of Multilingual and Multimodal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers, volume 4730 of Lecture Notes in Computer Science*, pages 237–248, 2007.

[11] A. Peñas, Á. Rodrigo, V. Sama, and F. Verdejo. Testing the reasoning for question answering validation. In *Journal of Logic and Computation. 18(3)*, pages 459–474, 2008.

[12] A. Peñas, Á. Rodrigo, and F. Verdejo. Overview of the Answer Validation Exercise 2007. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers, volume 5152 of Lecture Notes in Computer Science*, pages 237–248, 2007.

[13] J. Prager, E. Brown, A. Coden, and D. Radev. Question-answering by predictive annotation. In *Proceedings of the 23rd SIGIR Conference*, page 184–191, 2000.

[14] D. Sonntag. Distributed nlp and machine learning for question answering grid. In *Proceedings of the workshop on Semantic Intelligent Middleware for the Web and the Grid at the 16th European Conference on Artificial Intelligence (ECAI)*, 2004.