Overview of the Answer Validation Exercise 2008

Álvaro Rodrigo, Anselmo Peñas, Felisa Verdejo

Dpto. Lenguajes y Sistemas Informáticos, UNED
{alvarory, anselmo, felisa}@lsi.uned.es

Abstract

The Answer Validation Exercise at the Cross Language Evaluation Forum (CLEF) is aimed at developing systems able to decide whether the answer of a Question Answering (QA) system is correct or not. We present here the exercise description, the changes in the evaluation with respect to the last edition, and the results of this third edition (AVE 2008). Last year's changes allowed us to measure the possible gain in performance obtained by using AV systems as the selection method of QA systems. In this edition we wanted to reward AV systems able to detect if all the candidate answers to a question are incorrect. 9 groups have participated with 24 runs in 5 different languages, and compared with the QA systems, the results show an evidence of the potential gain that more sophisticated AV modules might introduce in the task of QA.

Keywords

Question Answering, Evaluation, Textual Entailment, Answer Validation

1. Introduction

The first Answer Validation Exercise (AVE 2006) [6] was activated two years ago in order to promote the development and evaluation of subsystems aimed at validating the correctness of the answers given by Question Answering (QA) systems. In some sense, systems must emulate human assessment of QA responses and decide whether an answer is correct or not according to a given supporting text. This automatic Answer Validation is expected to be useful for improving QA systems performance [4]. However, the evaluation methodology in AVE 2006 did not permit to quantify this improvement and thus, the exercise was modified in AVE 2007 [8], where the problem of Automatic Hypothesis Generation was also opened.

In AVE 2007 participant systems had to emulate QA systems selecting one answer per question from a set of candidate ones. These candidate answers were the ones given by QA systems participating at the QA main track at CLEF. This allowed us to study the use of Answer Validation (AV) systems as the answer selection method used by a QA system. Nevertheless, it was not acknowledged the ability of an AV system detecting if all the candidate answers to a question were incorrect. Systems with this ability could ask for new answers to the QA systems, opening the possibility of obtaining a correct answer to the question. Besides, NIL answers could be detected. Then, we have studied this behaviour in AVE 2008.

2. Exercise Description

Following the format proposed in AVE 2007, in this edition participant systems received a set of triplets (Question, Answer, Supporting Text) and they had to return a value for each triplet rejecting or accepting it. More in detail, the input format was a set of pairs (Answer, Supporting Text) grouped by Question (see Figure 1 for an example). Systems must consider the Question and validate each of these (Answer, Supporting Text) pairs. The number of answers to be validated per question depends on the number of participant systems at the QA main track.

```
<q id="0001" lang="EN">
        <q_str>What was the nationality of Jacques
        Offenbach?</q_str>
        <a id="0001_1" value="">
                <a_str>Germany</a_str>
                <t_str doc="Offenbach">Offenbach Offenbach Offenbach
                can refer to: The city Offenbach in Hesse,
                Germany.</t_str>
        </a>
        <a id="0001_2" value="">
                <a_str>France</a_str>
                <t_str doc="Jacques Offenbach">His son received the
                name "Jakob Offenbach" at birth, though he changed it
                to Jacques when he settled in France.</t_str>
        </a>
        <a id="0001_3" value="">
                <a_str>Thousand Oaks</a_str>
                <t_str doc="LA111794-0288">Ventura College's
                production of George Bernard Shaw's "Arms and the
                Man" and  Moorpark College's version of the Jacques
                Offenbach operetta "La Vie  Parisienne" are the
                costume shows; in Thousand Oaks, Cal Lutheran
                University is  mounting the contemporary drama "Minor
                Demons."</t_str>
        </a>
        ...
</q>
```

Figure 1. Excerpt of the English test collection in AVE 2008

Participant systems must return one of the following values for each answer according to the response format (see Figure 2):

- VALIDATED indicates that the answer is correct and supported by the given supporting text. There is no restriction in the number of VALIDATED answers returned per question (from zero to all).

- SELECTED indicates that the answer is VALIDATED and it is the one chosen as the output to the current question by a hypothetical QA system. The SELECTED answers are evaluated against the QA systems of the Main Track. No more than one answer per question can be marked as SELECTED. At least one of the VALIDATED answers must be marked as SELECTED.

- REJECTED indicates that the answer is incorrect or there is not enough evidence of its correctness. There is no restriction in the number of REJECTED answers per question (from zero to all).

```
q_id a_id [SELECTED|VALIDATED|REJECTED] confidence
```

Figure 2. Response format in AVE 2008

This configuration permitted us to compare the AV systems responses with the QA ones, and to obtain some evidences about the gain in performance that sophisticated AV modules can give to QA systems (see below).

3. Collections

In the exercise we want to promote the development of AV systems that perform an analysis beyond the use of redundancies in answers. Since the fact of grouping all the answers to the same question could lead to provide extra information based on counting answer redundancies, like in AVE 2007, if an answer is contained in another answer, we remove the shorter one. Besides, with this processing no extra information is given with respect to QA participant systems at the main track, allowing the comparison with them. Furthermore, NIL and void answers were discarded for building the collections. This processing lead to a reduction in the number of

answers initially available in the collections (see Tables 1 and 2): from 13.79% in the Italian development collection to 78.57% in the Bulgarian test collection.

Like in the past edition of QA@CLEF [3], questions were grouped by topic. In this organization by topics, the first question of each topic is self contained in the sense that there is no need of information outside the question to answer it. However, the rest of the topic questions can refer to implicit information linked to the previous questions and answers of the topic group (anaphora, co-reference, etc.). Therefore, for the AVE 2008 test collections we only made use of the self-contained questions (the first one of each topic group) and their respective answers given by the participant systems in QA.

For the assessments, we reused the QA judgements because they were done considering the supporting snippets in a similar way the AV systems must do. The relation between QA assessments and AVE judgements was the following:

- Answers judged as Correct in QA have a value equal to VALIDATED in AVE
- Answers judged as Wrong or Unsupported in QA have a value equal to REJECTED in AVE
- Answers judged as Inexact in QA have a value equal to UNKNOWN in AVE and are ignored for evaluation purposes.
- Answers not evaluated at the QA main track (if any) are also tagged as UNKNOWN in AVE and they are also ignored in the evaluation.

3.1 Development Collections

Development collections were obtained from the QA@CLEF 2006 [5] and 2007 [3] main track questions and answers. Table 1 shows the number of questions and answers for each language together with the percentage that these answers represent over the number of answers initially available, and the number of answers with VALIDATED and REJECTED values.

These collections were available for participants after their registration at CLEF at http://nlp.uned.es/clef-qa/ave/

	German	**English**	**Spanish**	**French**	**Italian**	**Dutch**	**Portuguese**	**Romanian**
Questions	108	67	169	118	100	78	148	82
Answers (final)	264	195	551	171	100	196	346	103
% over available answers	45.52%	58.21%	64.82%	68.95%	86.21%	50.26%	28.83%	42.21%
VALIDATED	67	21	127	85	16	31	148	45
REJECTED	197	174	424	86	84	165	198	58

Table 1. Number of questions and answers in the AVE 2008 development collections

3.2 Test Collections

Test collections were obtained from the runs sent to QA@CLEF 2008 main track [2]. In this edition, there were runs in 9 languages: German, English, Spanish, French, Bulgarian, Dutch, Portuguese, Romanian and Basque. Thus, a test collection in AVE was generated for each of these languages.

Table 2 shows the number of questions and the number of answers to be validated (or rejected) in the test collections together with the percentage that these answers represent over the answers initially available.

	German	English	Spanish	French	Bulgarian	Dutch	Portuguese	Romanian	Basque
Questions	119	160	136	108	27	128	149	119	104
Answers (final)	1027	1055	1528	199	27	228	1014	497	541
% over available answers	39.61%	57.37%	49.98%	60.30%	21.43%	42.54%	43.63%	48.58%	55.09%
VALIDATED	111	79	153	52	12	44	208	52	39
REJECTED	854	940	1354	126	9	177	747	406	483
UNKNOWN	62	36	21	21	6	7	59	39	19

Table 2. Number of questions and answers in the AVE 2008 test collections

4. Evaluation

In order to evaluate systems' performance, we used two groups of measures. In [7] was argued why the AVE evaluation is based on the detection of correct answers. Then, instead of using an overall accuracy, the first group of measures is composed by precision (1), recall (2) and F-measure (3) (harmonic mean) over answers that must be VALIDATED (in this first group, when a participant system returns SELECTED to an answer, the answer is considered as VALIDATED).

Results can be compared between systems but always taking as reference the following baselines:

1. A system that accepts all answers (returns VALIDATED or SELECTED in 100% of cases)
2. A system that accepts 50% of the answers (random)

$$precision = \frac{|\det ected\_as\_SELECTED\_or\_VALIDATED|}{|predicted\_as\_SELECTED\_or\_VALIDATED|} \quad (1)$$

$$recall = \frac{|\det ected\_as\_SELECTED\_or\_VALIDATED|}{|CORRECT\_answers|} \quad (2)$$

$$F = \frac{2 * recall * precision}{recall + precision} \quad (3)$$

The aim of this group of measures is to evaluate the performance of an AV system used for ranking or filtering answers. Nevertheless, this is an intrinsic evaluation that is not enough for comparing AVE results with QA results in order to obtain some evidence about the goodness of incorporating more sophisticated validation systems into QA architectures. Our aim was to obtain evidences of this improvement in a comparative and shared evaluation.

Then, the second group of measures aims at comparing QA systems performance with the potential gain that AV systems could add to them. The first of these measures is qa_accuracy (4), which was already used in AVE 2007. Since answers were grouped by questions and AV systems were requested to SELECT one or none of them, the resulting behaviour is comparable to a QA system: for each question there is no more than one SELECTED answer. The proportion of correctly selected answers is a measure comparable to the accuracy used in the QA Main Track and, therefore, we can compare AV systems taking as reference the QA systems performance over the questions involved in AVE test collections.

This measure has an upper bound given by the proportion of questions that have at least one correct answer (in its corresponding group). This upper bound corresponds to a perfect selection of the correct answers given by

all the QA systems at the main track. 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 the following baseline system: a system that validates 100% of the answers and selects randomly one of them. Thus, this baseline can be seen as the average proportion of correct answers per question group. We called this baseline random_qa_accuracy (6). Moreover, another baseline can be also taken into account. Since a good AV system should be able to yield the best QA system, we will consider the best QA system of each language as a baseline.

$$qa\_accuracy = \frac{|answers\_SELECTED\_correctly|}{|questions|} \quad (4)$$

$$\%\_best\_combination = \frac{|answers\_SELECTED\_correctly|}{|questions\_with\_correct\_answers|} * 100 \quad (5)$$

$$random\_qa\_accuracy = \frac{1}{|questions|} \sum_{q \in questions} \frac{|correct\_answers\_of(q)|}{|answers\_of(q)|} \quad (6)$$

The problem of qa_accuracy is that it 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, so in this edition we wanted to acknowledge this ability. The justification of why to acknowledge the recognizing of questions without correct answers arises from the fact that a possible gain in performance could be obtained in these questions. 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.

Therefore, we proposed the use of qa_rej_accuracy (7), which acknowledges systems capable of detecting when all the answers to a question are incorrect. Then, with this measure and qa_accuracy, we can propose qa_accuracy_max (8). This measure represents a range with a lower bound expressed by qa_accuracy and an upper bound that adds to qa_accuracy the accuracy that would be obtained answering correctly all the questions accounted in qa_rej_accuracy.

An estimation of the value obtained in this range is given by estimated_qa_performance (9). This measure considers that the questions accounted by qa_rej_accuracy are answered with the accuracy given by qa_accuracy. Therefore, this measure acknowledges a higher precision of AV systems detecting incorrect answers.

$$qa\_rej\_accuracy = \frac{|questions\_REJECTED\_correctly|}{|questions|} \quad (7)$$

$$qa\_accuracy\_\max = qa\_accuracy + qa\_rej\_accuracy \quad (8)$$

$$estimated\_qa\_performance = qa\_accuracy + qa\_rej\_accuracy * qa\_accuracy \quad (9)$$

5. Results

Nine groups (the same number that in the last edition) have participated in five different languages (German, English, Spanish, French and Romanian) with 24 runs. Table 3 shows the participant groups and the number of runs they submitted per language. Again, English and Spanish were the most popular with 8 and 6 runs respectively.

Tables 5-9 in the appendix show the results of Precision, Recall and F measure over correct answers for all participant systems in each language. Results cannot be compared between languages since the number of answers to be validated and the proportion of the correct ones are different for each language (due to the real submission of QA systems). However, they can be compared in each language with two baselines values that are given: the results of a system that always accepts all answers (validates 100% of the answers), and the results of a hypothetical system that validates the 50% of answers.

	German	English	Spanish	French	Romanian	Total
Fernuniversität in Hagen (FUH)	2					**2**
LIMSI				2		**2**
U. Iasi		2			2	**4**
DFKI	1	1				**2**
INAOE			2			**2**
U. Alicante		1	2			**3**
UNC		2				**2**
U. Jaén (UJA)		2	2	2		**6**
LINA				1		**1**
Total	**3**	**8**	**6**	**5**	**2**	**24**

Table 3. Participants and runs per language in AVE 2008

In our opinion, F-measure is an appropriate measure to identify the systems that perform better, measuring their ability to detect the correct answers and only them. However, it is also important to try to obtain some evidences about the improvement that AV systems could provide to QA systems. Tables 10-14 in the appendix show the rankings of systems (merging QA and AV systems) according to estimated_qa_performance calculated only over the subset of questions considered in AVE 2008. The tables contain also the information about the results of QA and AVE systems using the measures qa_accuracy, %_best_combination, qa_rej_accuracy and qa_accuracy_max. The values of qa_accuracy and estimated_qa_performance are the same in QA systems. Again, results cannot be compared between different languages, but they can be compared with the random baselines and with the results of the best QA system (which is marked with a shadow).

The graphic interpretations of these tables are shown in Figures 3-7 in the appendix. In these graphics the value of qa_accuracy is 1 in the perfect selection baseline. This corresponds to a perfect selection of a correct answer (if any) per question and the detection of all the questions with no correct answers (qa_rej_accuracy). However, the value of estimated_qa_performance in this baseline is not 1 because it is assumed that the questions detected in qa_rej_accuracy will be answered with a precision value equal to the qa_accuracy of the perfect selection baseline. This value represents the accuracy of the best combination of the QA systems involved, which is not perfect.

5.1. Analysis of results

In three languages (German, English and Romanian) there has been at least one AV system performing better than the best QA system. In the languages where the best value of qa_accuracy was not obtained by an AV system, the best QA system outperforms in more than a 50% the following QA systems. If we see an AV system as a multi-stream selector of candidate answers, then AV systems follow a behaviour similar to an ensemble of classifiers. In Machine Learning (ML), an ensemble of classifiers is likely to be more accurate than an individual classifier except in the case of an element of the assemble outperforms in a high percent the rest of the classifiers

[1]. Therefore, it seems obvious that there must be more work focused in performing a better selection in this kind of situations.

5.2. Analysis of the measures

Regarding the use of the new measure estimated_qa_performance, the rankings are very similar to the ones obtained ranking by qa_accuracy. In fact, there have been only two changes, which are located in the English ranking (see Table 13 in the appendix). Firstly, the system uaic_2 obtains a better performance than ofe according to qa_accuracy (0.24 against 0.19). However, according to estimated_qa_performance, ofe is better than uaic_2 (0.27 against 0.24). This means that uaic_2 is better selecting correct answers. Nevertheless, if we consider the possible gain in performance that might be obtained detecting that all the answers to a question are incorrect and asking for new ones to the QA systems, then ofe is better. Therefore, the system ofe may help to obtain better results in QA than the system uaic_2. Besides, it can be seen how the ranking according to estimated_qa_performance is more similar to the one given by F-measure, which in some way, also considers the precision of a system detecting incorrect answers.

The second change in the rankings involves the QA system dfki081deen, which has a better performance than the AVE system jota_2 according to qa_accuracy. However, according to estimated_qa_performance, the two systems have the same performance. Again, this indicates that AV systems detecting incorrect answers could lead to a better performance in QA.

Thus, it seams that estimated_qa_performance is a better measure for AV systems than qa_accuracy because it takes into account the ability of a system rejecting incorrect answers. Thus, it is given a better estimation of the performance obtained by using AV systems in QA. Furthermore, the rankings are more similar to the ones obtained by using F-measure.

5.3. Analysis of the techniques used

All the participants have reported the use of textual entailment in their systems except two groups (LINA and LIMSI). However, while in the past edition the half of the participants reported the use of automatic hypothesis generation, in this edition only two participants (U. Iasi and U. Alicante) have used it. 6 of the 9 groups (FUH, U. Iasi, INAOE, DFKI, U. Alicante and LIMSI) have also participated in the QA main track, showing that there is a growing interest in using AV in QA participant systems at CLEF.

Table 4 shows the techniques used by AVE participant systems. Following the tendency showed in the past edition, all the systems have reported the use of lexical processing. Moreover, this year there are more groups using syntactic processing, mainly chunking or dependency analysis. Except in Spanish, where none system reported the use of syntactic processing, the system with the best result in each language performed some kind of syntactic processing, mainly by means of dependency parsing. However, the use of semantic analysis has decreased while the use of WordNet has been increased (50% of participants used it). Furthermore, there has been a high increase in the use of Named Entities, with 7 of 9 groups considering them. Therefore, it seems that it can be an important information to be taken into account in AV.

All the participants except two systems (U. Iasi and LINA) have used ML for taking the validation decision, following the tendency of the last edition. Besides, ML was used by the participants with the best score in each language. While lexical similarity was the most common feature used, syntactic similarity was included by the half of the participants. However, semantics features were taken into account by very few participants. Only one participant (FUH) reported the use of a theorem prover this year. Support vector machines (SVM) and decision trees were the most used classifiers. Nevertheless, there are not evidences about the best performance of one or another of these classifiers.

Finally, after a comparison between the resources taken into account and the results obtained, it seems that more resources do not imply better performance. In fact, systems performing semantic analysis have not achieved the best results in their languages.

	U. Iasi	INAOE	FUH	DFKI	UJA	U. Alicante	LIMSI	LINA	UNC
Generates hypotheses	X					X			
Wordnet	X		X		X	X			X
Chunking	X				X		X	X	
n-grams, longest common Subsequences				X	X	X			X
Phrase transformations	X						X		
NER	X	X	X	X		X	X	X	
Num. expressions	X	X	X	X		X	X	X	
Temp. expressions	X	X	X				X	X	
Coreference resolution									
Dependency analysis	X			X			X		
Syntactic similarity									
Functions (sub, obj, etc)	X			X			X		
Syntactic transformations	X						X		
Word-sense disambiguation			X						
Semantic parsing	X		X						
Semantic role labeling			X						
First order logic representation	X		X						
Theorem prover			X						
Semantic similarity									

Table 4. Information about the techniques used by the AVE participants.

6. Conclusions

In AVE 2008 there has been the same number of participants of last year (9) in 5 different languages. However, 8 more runs have been sent, showing a growing interest in the task.

Results show that AV systems could improve the performance of current QA systems. This improvement comes when AV systems are used for selecting the final answer from a set of candidate ones. In fact, according to the results, except in the languages where the best QA system outperforms the others QA systems in more than a 50%, there was an AV system with better performance than QA systems.

In this edition new measures have been introduced in order to obtain a more informative estimation of the potential of AV systems in QA performance. These new measures reward the ability of some systems detecting if all the candidate answers to a question are incorrect. These measures have shown to be very useful when two systems have a similar performance according to qa_accuracy. In this situation, the new measure estimated_qa_performance have indicated that AV systems with a better precision detecting incorrect answers would be more useful in QA because more answers could be asked to QA systems when all the candidate answers to a question are incorrect. Then, a correct answer might be found.

The most used technique continues being lexical processing while the use of syntactic analysis has grown. Nevertheless, very few systems have performed semantic analysis. Besides, a high percent of participants have combined different features using ML. Finally, the best systems performed both lexical and syntactic analysis, and they consider NE.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Innovation within the project QEAVis-Catiex (TIN2007-67581-C02-01), the Education Council of the Regional Government of Madrid and the European Social Fund. We are grateful to all the people involved in the organization of the QA track (specially to the coordinators at CELCT, Danilo Giampiccolo and Pamela Forner).

References

1. T. Dietterich. Machine-learning research: Four current directions. AI Magazine, pages 97-136, Winter 1997.
2. D. Giampiccolo et al. Overview of the CLEF 2008 Multilingual Question Answering Track. Working Notes of CLEF 2008. 2008.
3. D. Giampiccolo, P. Forner, J. Herrera, A. Peñas, C. Ayache, D. Cristea, V. Jijkoun, P. Osenova, P. Rocha, B. Sacaleanu, R. Sutcliffe. Overview of the CLEF 2007 Multilingual Question Answering Track. CLEF 2007, Lecture Notes in Computer Science LNCS 5152. Springer. Berlin. 2008.
4. S. Harabagiu, 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, pages 905-912, Sydney. 2006.
5. 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. CLEF 2006, Lecture Notes in Computer Science LNCS 4730. Springer. Berlin. 2007.
6. A. Peñas, Á. Rodrigo, V. Sama, F. Verdejo. Overview of the Answer Validation Exercise 2006. CLEF 2006, Lecture Notes in Computer Science LNCS 4730. Springer, Berlin. 2007.
7. A. Peñas, Á. Rodrigo, V. Sama, F. Verdejo. Testing the Reasoning for Question Answering Validation. Journal of Logic and Computation 2007.
8. A. Peñas, Á. Rodrigo, F. Verdejo. Overview of the Answer Validation Exercise 2007. CLEF 2007, Lecture Notes in Computer Science LNCS 5152. Springer, Berlin. 2008.

Appendix

The following tables show the values of Precision, Recall and F measure over correct answers of AVE participant systems in different languages.

Group	System	F	Precision	Recall
DFKI	ltqa	0,61	0,54	0,71
FUH	glockner_1	0,39	0,33	0,49
FUH	glockner_2	0,29	0,25	0,34
100% VALIDATED		0,21	0,12	1
50% VALIDATED		0,19	0,12	0,5

Table 5. Precision, Recall and F measure over correct answers for German

Group	System	F	Precision	Recall
UA	ofe_2	0,44	0,32	0,67
INAOE	tellez_2	0,39	0,30	0,59
UA	ofe_1	0,38	0,26	0,76
INAOE	tellez_1	0,23	0,13	0,86
100% VALIDATED		0,18	0,10	1
50% VALIDATED		0,17	0,10	0,5
UJA	magc_1(timbl)	0,06	0,15	0,04
UJA	magc_2(bbr)	0,05	0,22	0,03

Table 6. Precision, Recall and F measure over correct answers for Spanish

Group	System	F	Precision	Recall
LIMSI	bgrau_1	0,61	0,75	0,52
LIMSI	bgrau_2	0,57	0,88	0,42
LINA	monceaux	0,51	0,56	0,46
100% VALIDATED		0,45	0,29	1
50% VALIDATED		0,37	0,29	0,5
UJA	magc_1(timbl)	0,08	0,15	0,06
UJA	magc_2(bbr)	0,08	0,13	0,06

Table 7. Precision, Recall and F measure over correct answers for French

Group	System	F	Precision	Recall
DFKI	ltqa	0,64	0,54	0,78
UA	ofe	0,49	0,35	0,86
UNC	jota_2	0,21	0,13	0,56
Iasi	uaic_2	0,19	0,11	0,85
UNC	jota_1	0,17	0,09	0,94
Iasi	uaic_1	0,17	0,09	0,76
100% VALIDATED		0,14	0,08	1
50% VALIDATED		0,13	0,08	0,5
UJA	magc_2(bbr)	0,02	0,17	0,01
UJA	magc_1(timbl)	0	0	0

Table 8. Precision, Recall and F measure over correct answers for English.

Group	System	F	Precision	Recall
Iasi	uaic_2	0,23	0,13	0,92
Iasi	uaic_1	0,22	0,12	0,92
100% VALIDATED		0,20	0,11	1
50% VALIDATED		0,19	0,11	0,50

Table 9. Precision, Recall and F measure over correct answers for Romanian.

The following tables and graphics show the comparison of AV systems performance with QA systems of AVE participant systems in different languages.

System	System type	estimated_ qa_performance	qa_accuracy (% best combination)	qa_rej_ accuracy	qa_ accuracy_max
Perfect selection		0,77	0,52 (100%)	0,48	1
ltqa	AV	0,52	0,43 (82,26%)	0,21	0,64
dfki082dede	QA	0,38	0,38 (72,58%)	0	0,38
dfki081dede	QA	0,37	0,37 (70,97%)	0	0,37
glockner_1	AV	0,32	0,32 (61,29%)	0	0,32
fuha082dede	QA	0,24	0,24 (45,16%)	0	0,24
glockner_2	AV	0,23	0,23 (43,55%)	0	0,23
fuha081dede	QA	0,22	0,22 (41,94%)	0	0,22
loga081dede	QA	0,17	0,17 (32,26%)	0	0,17
fuha082ende	QA	0,16	0,16 (30,65%)	0	0,16
fuha081ende	QA	0,16	0,16 (30,65%)	0	0,16
loga082dede	QA	0,15	0,15 (29,03%)	0	0,15
dfki081ende	QA	0,14	0,14 (27,42%)	0	0,14
fuha081esde	QA	0,12	0,12 (22,58%)	0	0,12
Random		0.11	0.11 (21.13%)	0	0.11
fuha082esde	QA	0,10	0,10 (19,35%)	0	0,10

Table 10. Comparing AV systems performance with QA systems in German

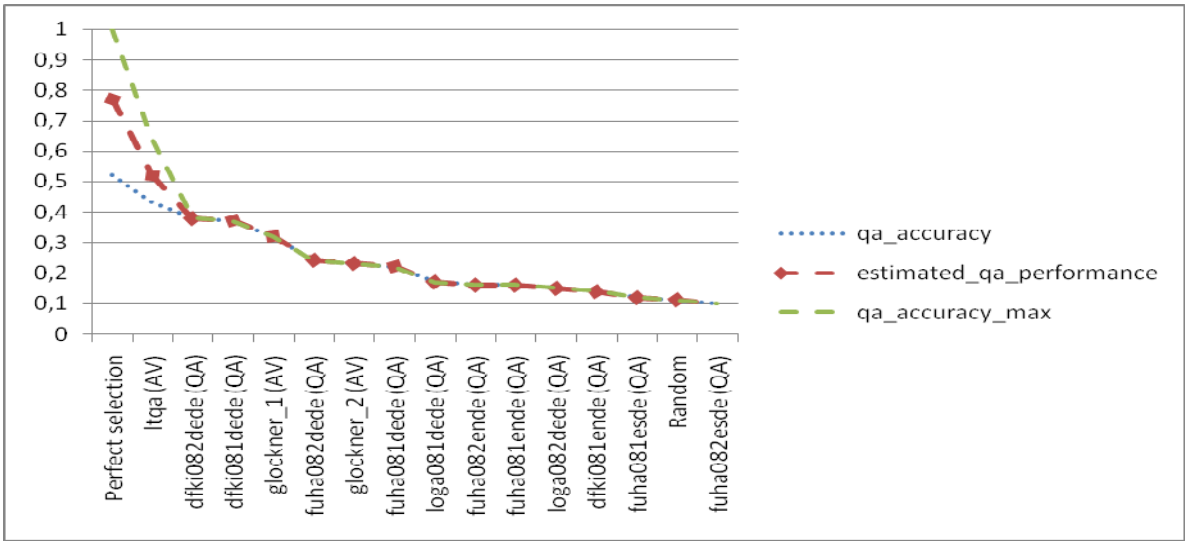

Figure 3. Graphic comparing AV systems performance with QA systems in German

System	System type	estimated_ qa_performance	qa_accuracy (% best combination)	qa_rej_ accuracy	qa_ accuracy_max
Perfect selection		0,85	0,62 (100%)	0,38	1
prib081eses	QA	0,54	0,54 (88,10%)	0	0,54
ofe_1	AV	0,37	0,32 (52,38%)	0,14	0,46
tellez_1	AV	0,34	0,32 (52,38%)	0,06	0,38
ofe_2	AV	0,33	0,27 (44,05%)	0,21	0,48
tellez_2	AV	0,33	0,27 (44,05%)	0,22	0,49
inao081eses	QA	0,25	0,25 (40,48%)	0	0,25
inao082eses	QA	0,25	0,25 (40,48%)	0	0,25
qaua082eses	QA	0,22	0,22 (35,71%)	0	0,22
mira081eses	QA	0,21	0,21 (33,33%)	0	0,21
mira082eses	QA	0,18	0,18 (29,76%)	0	0,18
qaua081enes	QA	0,18	0,18 (28,57%)	0	0,18
qaua082enes	QA	0,13	0,13 (21,43%)	0	0,13
qaua081eses	QA	0,12	0,12 (19,05%)	0	0,12
Random		0,11	0,11 (17,12%)	0	0,11
mira081fres	QA	0,06	0,06 (9,52%)	0	0,06
magc_1(timbl)	AV	0,06	0,04 (7,14%)	0,32	0,36
magc_2(bbr)	AV	0,03	0,02 (3,57%)	0,35	0,37

Table 11. Comparing AV systems performance with QA systems in Spanish

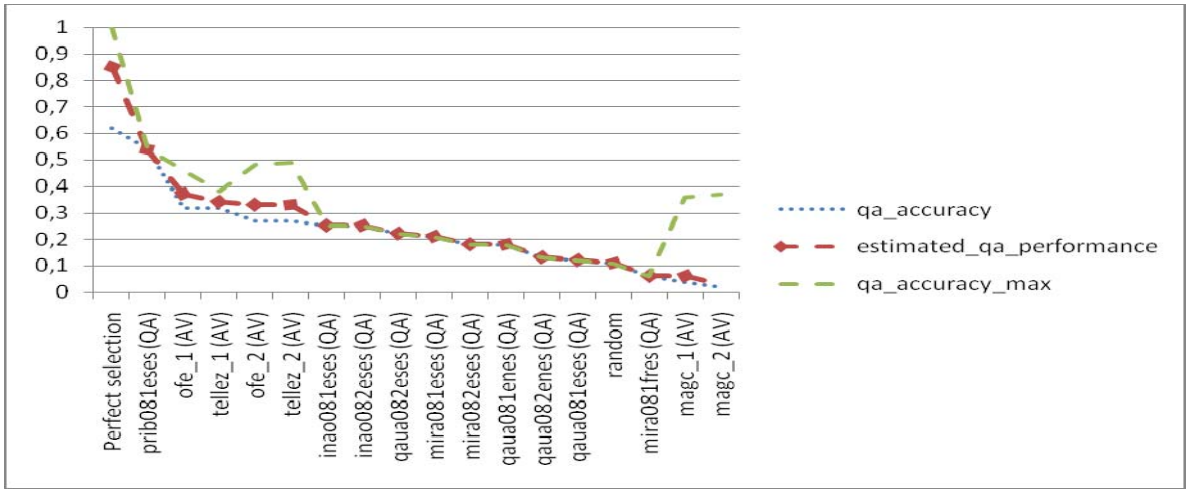

Figure 4. Graphic comparing AV systems performance with QA systems in Spanish

System	System type	estimated_ qa_performance	qa_accuracy	qa_rej_ accuracy	qa_ accuracy_max
Perfect selection		0,73	0,48 (100%)	0,52	1
syna081frfr	QA	0,47	0,47 (98,08%)	0	0,47
Random		0,33	0,33 (68,80%)	0	0,33
bgrau_1	AV	0,32	0,23 (48,08%)	0,39	0,62
monceaux	AV	0,29	0,21 (44,23%)	0,35	0,56
bgrau_2	AV	0,29	0,19 (40,38%)	0,48	0,67
syna081ptfr	QA	0,19	0,19 (40,38%)	0	0,19
syna081enfr	QA	0,17	0,17 (34,62%)	0	0,17
magc_1(timbl)	AV	0,04	0,03 (5,77%)	0,41	0,44
magc_2(bbr)	AV	0,04	0,03 (5,77%)	0,41	0,44

Table 12. Comparing AV systems performance with QA systems in French

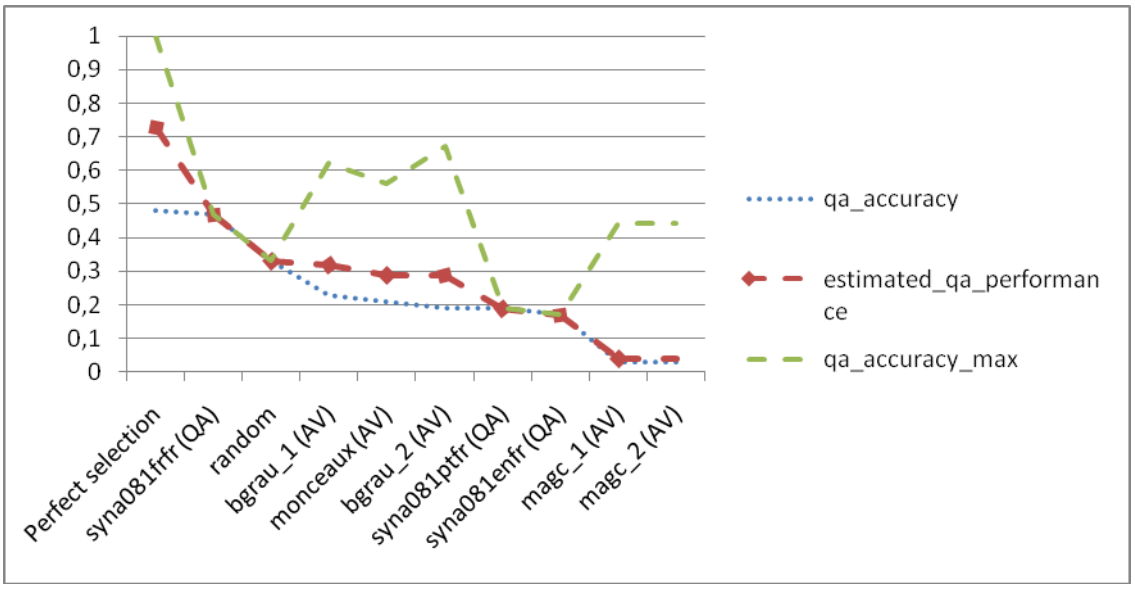

Figure 5. Graphic comparing AV systems performance with QA systems in French

System	System type	estimated_ qa_performance	qa_accuracy (% best combination)	qa_rej_ accuracy	qa_ accuracy_max
Perfect selection		0,56	0,34 (100%)	0,66	1
ltqa	AV	0,34	0,24 (70,37%)	0,44	0,68
ofe	AV	0,27	0,19 (57,41%)	0,4	0,59
uaic_2	AV	0,24	0,24 (70,37%)	0,01	0,25
wlvs081roen	QA	0,21	0,21 (62,96%)	0	0,21
uaic_1	AV	0,19	0,19 (57,41%)	0	0,19
jota_2	AV	0,17	0,16 (46,30%)	0,1	0,26
dfki081deen	QA	0,17	0,17 (50%)	0	0,17
jota_1	AV	0,16	0,16 (46,30%)	0	0,16
dcun081deen	QA	0,10	0,10 (29,63%)	0	0,10
Random		0,09	0,09 (25,25%)	0	0,09
nlel081enen	QA	0,06	0,06 (18,52%)	0	0,06
nlel082enen	QA	0,05	0,05 (14,81%)	0	0,05
ilkm081nlen	QA	0,04	0,04 (12,96%)	0	0,04
magc_2(bbr)	AV	0,01	0,01 (1,85%)	0,64	0,65
dcun082deen	QA	0,01	0,01 (1,85%)	0	0,01
magc_1(timbl)	AV	0	0 (0%)	0,63	0,63

Table 13. Comparing AV systems performance with QA systems in English

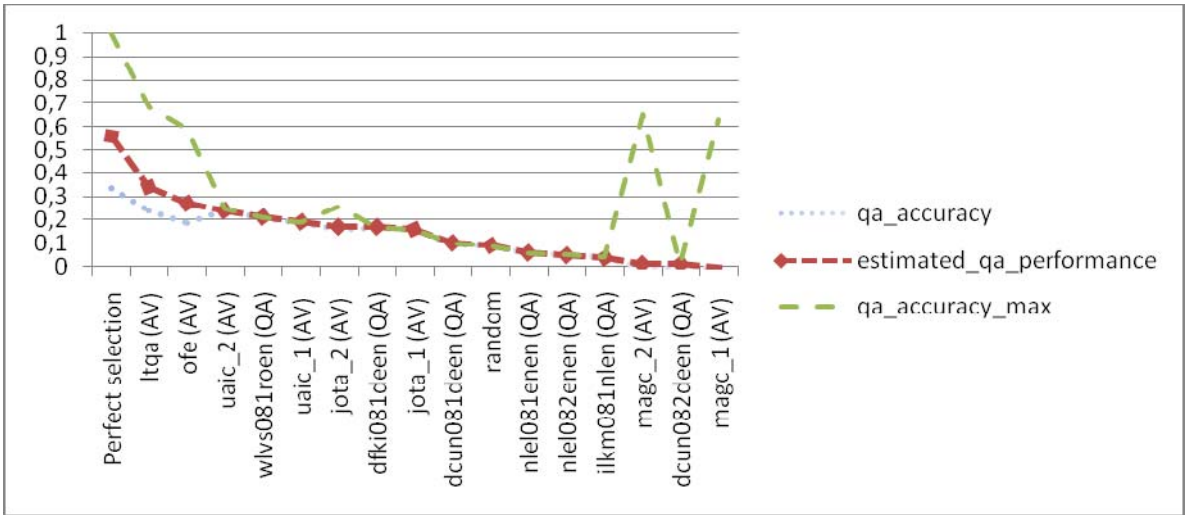

Figure 6. Graphic comparing AV systems performance with QA systems in English

System	System type	estimated_ qa_performance	qa_accuracy (% best combination)	qa_rej_ accuracy	qa_ accuracy_max
Perfect selection		0,65	0,41 (100%)	0,59	1
uaic_2	AV	0,25	0,24 (57,14%)	0,05	0,29
UAIC082roro	QA	0,22	0,22 (53,06%)	0	0,22
UAIC081roro	QA	0,19	0,19 (46,94%)	0	0,19
uaic_1	AV	0,17	0,17 (40,82%)	0	0,17
icia082roro	QA	0,17	0,17 (40,82%)	0	0,17
Random		0,10	0,10 (24,66%)	0	0,10
icia081roro	QA	0,08	0,08 (18,37%)	0	0,08

Table 14. Comparing AV systems performance with QA systems in Romanian

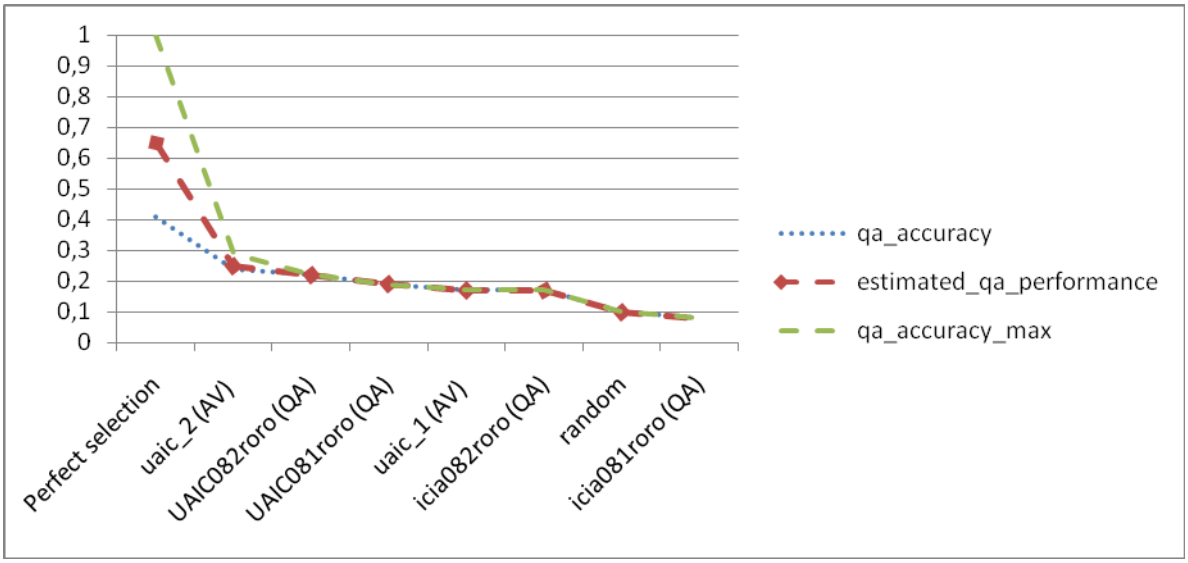

Figure 7. Graphic comparing AV systems performance with QA systems in Romanian