Overview of the Answer Validation Exercise 2006

Anselmo Peñas, Álvaro Rodrigo, Valentín Sama, Felisa Verdejo

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

Abstract

The first Answer Validation Exercise (AVE) has been launched at the Cross Language Evaluation Forum 2006. This task is aimed at developing systems able to decide whether the answer of a Question Answering system is correct or not. The exercise is described here together with the evaluation methodology and the systems results. The starting point for the AVE 2006 was the reformulation of the Answer Validation as a Recognizing Textual Entailment problem, under the assumption that hypothesis can be automatically generated instantiating hypothesis patterns with the QA systems' answers. 11 groups have participated with 38 runs in 7 different languages. Systems that reported the use of logic have obtained the best results in their respective subtasks.

Keywords

Question Answering, Evaluation, Textual Entailment, Answer Validation

1. Introduction

The first Answer Validation Exercise (AVE 2006) was activated to promote the development and evaluation of subsystems aimed at validating the correctness of the answers given by QA systems. This automatic Answer Validation is expected to be useful for improving QA systems performance, help humans in the assessment of QA systems output, improve systems confidence self-score, and to develop better criteria for collaborative systems.

Systems must emulate human assessment of QA responses and decide whether an answer is correct or not according to a given snippet. The first AVE has been reformulated as Textual Entailment problem [1][2] where the hypotheses have been built semi-automatically turning the questions plus the answers into an affirmative form.

Participant systems received a set of pairs text-hypothesis built from the QA main track responses of the CLEF 2006, following the methodology described in [6]. Development collections were built from the QA assessments of last campaigns [3][4][5][7] in English and Spanish. A subtask per language has been activated: English, Spanish, French, German, Dutch, Italian, Portuguese and Bulgarian.

Participant systems must return a value YES or NO for each pair text-hypothesis to indicate if the text entails the hypothesis or not (i.e. the answer is correct according to the text). Systems results are evaluated against the QA human assessments.

The training collections together with the 8 testing collections (one per language) resulting from the first AVE 2006 are available at http://nlp.uned.es/QA/ave for researchers registered at CLEF.

Section 2 describe the test collections. Section 3 motivates the evaluation measures. Section 4 presents the results in each language and Section 5 present some conclusions and future work.

2. Test Collections

As a difference with the previous campaigns of the QA track, a text snippet was requested to support the correctness of the answers. The QA assessments were done considering the given snippet, so the direct relation between QA assessments and RTE judges was preserved: Pairs corresponding to answers judged as *Correct* have an entailment value equal to YES; pairs corresponding to answers judged as *Wrong* or *Unsupported* have an entailment value equal to NO; and pairs corresponding to answers judged as *Inexact* have an entailment value equal to UNKNOWN and are ignored for evaluation purposes. Pairs coming from answers not evaluated at the QA Track are also tagged as UNKNOWN and they are also ignored in the evaluation.

Figure 1 resumes the process followed in each language to build the test collection. Starting with the 200 questions, a hypothesis pattern was created for each one, and instantiated with all the answers of all systems for the corresponding question. The pairs were completed with the text snippet given by the system for supporting the answer.

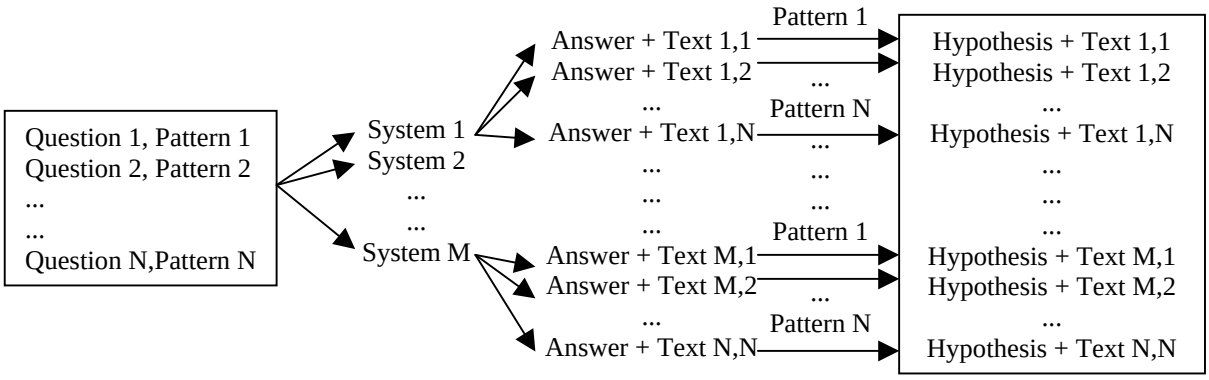

Figure 1. Text-hypothesis pairs for the Answer Validation Exercise from the pool of answers of the main QA Track.

Table 1 shows the number of pairs for each language obtained as the result of the processing. This pairs conform the test collections for each language and a benchmark for future evaluations.

Table 1. YES, NO and UNKNOWN pairs in the testing collections of AVE 2006

	German	**English**	**Spanish**	**French**	**Italian**	**Dutch**	**Portuguese**
YES pairs	344(24%)	198(9.5%)	671(28%)	705(22%)	187(16%)	81(10%)	188(14%)
NO pairs	1064(74%)	1048(50%)	1615(68%)	2359(72%)	901(79%)	696(86%)	604(46%)
UNKNOWN	35(3%)	842(40.5%)	83(4%)	202(6%)	52(5%)	30(4%)	532(40%)
Total	1443	2088	2369	3266	1140	807	1324

Percentages of YES, NO and UNKNOWN pairs are similar in all languages except for the percentage of UNKNOWN pairs in English and Portuguese, in which up to 5 runs weren't finally assessed in the QA task and therefore, the corresponding pairs couldn't be used to evaluate the systems.

3. Evaluation of the Answer Validation Exercise

The evaluation is based on the detection of the correct answers and only them. There are two reasons for this. First, an answer will be validated if there is enough evidence to affirm its correctness. *Figure 2* shows the decision flow that involves an Answer Validation module after searching for candidate answers: In the cases where there is not enough evidence of correctness (according to the AV module), the system must request another candidate answer. Thus, the Answer Validation must focus on detecting that there is enough evidence of the answer correctness.

Second, in a real exploitation environment, there is no balance between correct and incorrect candidate answers, that is to say, a system that validates QA responses does not receive correct and incorrect answers in the same proportion. In fact, the experiences at CLEF during the last years showed that only 23% of all the answers given by all the systems were correct (results for the Spanish as target, see [6]). Although numbers are expected to change, the important thing is that the evaluation of Answer Validation modules must consider the real output of Question Answering systems, which is not balanced. We think this leads to different development strategies closer to the real AV Exercise that, anyway, must be evaluated with this unbalanced nature.

Therefore, instead of using an overall accuracy as the evaluation measure, we proposed to use precision (1), recall (2) and a F-measure (3) (harmonic mean) over pairs with entailment value equals to YES. In other words, we proposed to quantify systems ability to detect the pairs with entailment or to detect whether there is enough evidence to accept an answer. If we would had considered the accuracy over all pairs then a baseline AV system that always answers NO (rejects all answers) would obtain an accuracy value of 0.77, which seems too high for evaluation purposes.

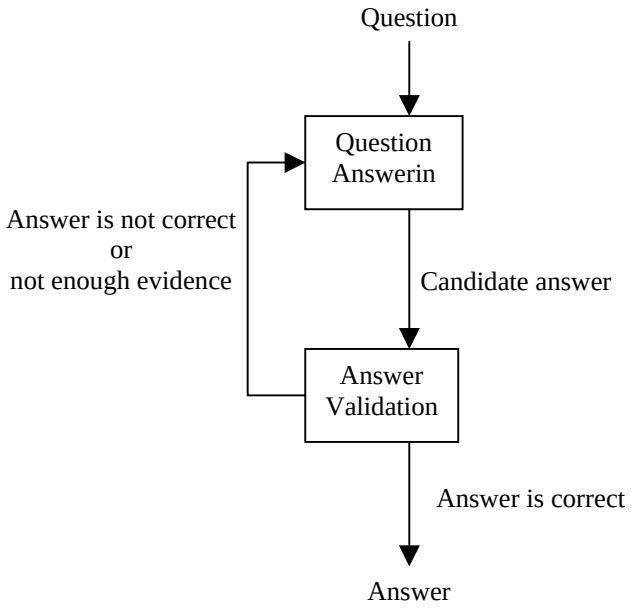

Figure 2. Decision flow for the Answer Validation

$$precision = \frac{|\, predicted\_as\_YES\_correcly \,|}{|\{predicted\_as\_YES\} \cap \{\overline{UNK\_pairs}\}|} \qquad (1)$$

$$recall = \frac{|\, predicted\_as\_YES\_correctly \,|}{|\, YES\_pairs \,|} \qquad (2)$$

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall} \qquad (3)$$

In the other hand, the higher the proportion of YES pairs is, the higher the baselines are. Thus, results can be compared between systems and always taking as reference the baseline of a system that accept all answers (return YES in 100% of cases). Since UNKNOWN pairs are ignored in the evaluation (though they were present in the test collection), the precision formula (2) was modify to ignore the cases were systems assessed a YES value to the UNKNOWN pairs.

4. Results

Eleven groups have participated in seven different languages at this first AVE 2006. Table 2 shows the participant groups and the number of runs they submitted per language. Al least two different groups participated

for each language, so the comparison between different approaches is possible. English and Spanish were the most popular with 11 and 9 runs respectively.

Table 2. Participants and runs per language in AVE 2006

	German	English	Spanish	French	Italian	Dutch	Portuguese	Total
Fernuniversität in Hagen	2							2
Language Computer Corporation		1	1					2
U. Rome "Tor Vergata"		2						2
U. Alicante (Kozareva)	2	2	2	2	2	2	1	13
U. Politecnica de Valencia		1						1
U. Alicante (Ferrández)		2						2
LIMSI-CNRS				1				1
U. Twente	1	2	2	1	1	2	1	10
UNED (Herrera)			2					2
UNED (Rodrigo)			1					1
ITC-irst		1						1
R2D2 project			1					1
Total	5	11	9	4	3	4	2	38

Only 3 of the 12 groups (FUH, LCC and ITC-IRST) have participated in the Question Answering Track showing the chance for new-comers to start developing a single QA module and, at the same time, open a place for experienced groups in RTE and KR to apply their research to the QA problem. We expect that in a near future the QA systems will take advantage of this communities working in the kind of reasoning needed for the Answer Validation.

Tables 3-9 show the results for all participant system in each language. Since the number of pairs and the proportion of the YES pairs is different for each language (due to the real submission of the QA systems), results can't be compared between languages. Together with the systems precision, recall and F-measure, two baselines values are shown: the results of a system that always accept all answers (returns YES in 100% of the pairs), and the results of a hypothetical system that returns YES for the 50% of pairs.

In the languages where at least one system reported the use of Logic (Spanish, English and German) the best performing system was one of them. Although the use of Logic doesn't guarantee a good result, the best systems used it. However, the most extensively used techniques were Machine Learning and overlapping measures between text and hypothesis.

Table 3. AVE 2006 Results for English

System Id	Group	F-measure	Precision	Recall	Techniques
COGEX	LCC	0.4559	0.3261	0.7576	Logic
ZNZ – TV_2	U. Rome	0.4106	0.2838	0.7424	ML
itc-irst	ITC-irst	0.3919	0.3090	0.5354	Lexical, Syntax, Corpus, ML
ZNZ – TV_1	U. Rome	0.3780	0.2707	0.6263	ML
MLEnt_2	U. Alicante	0.3720	0.2487	0.7374	Overlap, Corpus, ML
uaofe_2	U. Alicante	0.3177	0.2040	0.7172	Lexical, Syntax, Logic
MLEnt_1	U. Alicante	0.3174	0.2114	0.6364	Overlap, Logic, ML
uaofe_1	U. Alicante	0.3070	0.2144	0.5404	Lexical, Syntax, Logic
utwente.ta	U. Twente	0.3022	0.3313	0.2778	Syntax, ML
utwente.lcs	U. Twente	0.2759	0.2692	0.2828	Overlap, Paraphrase
100% YES Baseline		0.2742	0.1589	1	
50% YES Baseline		0.2412	0.1589	0.5	
ebisbal	U.P. Valencia	0.075	0.2143	0.0455	ML

Table 4. AVE 2006 Results for French

System Id	Group	F-measure	Precision	Recall	Techniques
MLEnt_2	U. Alicante	0.4693	0.3444	0.7362	Overlap, ML
MLEnt_1	U. Alicante	0.4085	0.3836	0.4369	Overlap, Corpus, ML
100% YES Baseline		0.3741	0.2301	1	
50% YES Baseline		0.3152	0.2301	0.5	
LIRAVE	LIMSI-CNRS	0.1112	0.4327	0.0638	Lexical, Syntax, Paraphrase
utwente.lcs	U. Twente	0.0943	0.4625	0.0525	Overlap

Table 5. AVE 2006 Results for Spanish

System Id	Group	F-measure	Precision	Recall	Techniques
COGEX	LCC	0.6063	0.527	0.7139	Logic
UNED_1	UNED	0.5655	0.467	0.7168	Overlap, ML
UNED_2	UNED	0.5615	0.4652	0.7079	Overlap, ML
NED	UNED	0.5315	0.4364	0.6796	NE recognition
MLEnt_2	U. Alicante	0.5301	0.4065	0.7615	Overlap, ML
R2D2	R2D2 Project	0.4938	0.4387	0.5648	Voting, Overlap, ML
utwente.ta	U. Twente	0.4682	0.4811	0.4560	Syntax, ML
100% YES Baseline		0.4538	0.2935	1	
utwente.lcs	U. Twente	0.4326	0.5507	0.3562	Overlap, Paraphrase
MLEnt_1	U. Alicante	0.4303	0.4748	0.3934	Overlap, Corpus, ML
50% YES Baseline		0.3699	0.2935	0.5	

Table 6. AVE 2006 Results for German

System Id	Group	F measure	Precision	Recall	Techniques
FUH_1	Fernuniversität in Hagen	0.5420	0.5839	0.5058	Lexical, Syntax, Semantics, Logic, Corpus
FUH_2	Fernuniversität in Hagen	0.5029	0.7293	0.3837	Lexical, Syntax, Semantics, Logic, Corpus, Paraphrase
MLEnt_2	U. Alicante	0.4685	0.3573	0.6802	Overlap, ML
100% YES Baseline		0.3927	0.2443	1	
MLEnt_1	U. Alicante	0.3874	0.4006	0.375	Overlap, Corpus, ML
50% YES Baseline		0.3282	0.2443	0.5	
utwente.lcs	U. Twente	0.1432	0.4	0.0872	Overlap

Table 7. AVE 2006 Results for Dutch

System Id	Group	F measure	Precision	Recall	Techniques
utwente.ta	U. Twente	0.3871	0.2874	0.5926	Syntax, ML
MLEnt_1	U. Alicante	0.2957	0.189	0.6790	Overlap, Corpus, ML
MLEnt_2	U. Alicante	0.2548	0.1484	0.9012	Overlap, ML
utwente.lcs	U. Twente	0.2201	0.2	0.2469	Overlap, Paraphrase
100% YES Baseline		0.1887	0.1042	1	
50% YES Baseline		0.1725	0.1042	0.5	

Table 8. AVE 2006 Results for Portuguese

System Id	Group	F measure	Precision	Recall	Techniques
100% YES Baseline		0.3837	0.2374	1	
utwente.lcs	U. Twente	0.3542	0.5783	0.2553	Overlap
50% YES Baseline		0.3219	0.2374	0.5	
MLEnt	U. Alicante	0.1529	0.1904	0.1277	Corpus

Table 9. AVE 2006 Results for Italian

System Id	Group	F measure	Precision	Recall	Techniques
MLEnt_2	U. Alicante	0.4066	0.2830	0.7219	Overlap, ML
MLEnt_1	U. Alicante	0.3480	0.2164	0.8877	Overlap, Corpus, ML
100% YES Baseline		0.2934	0.1719	1	
50% YES Baseline		0.2558	0.1719	0.5	
utwente.lcs	U. Twente	0.1673	0.3281	0.1123	Overlap

5. Conclusions and future work

The starting point for the AVE 2006 was the reformulation of the Answer Validation as a Recognizing Textual Entailment problem, under the assumption that hypothesis can be automatically generated instantiating hypothesis patterns with the QA systems answers. Thus, the collections developed in AVE are specially oriented to the development and evaluation of Answer Validation systems. We have also proposed a methodology for the evaluation in chain with a QA Track.

11 groups have participated with 38 runs in 7 different languages. Systems that reported the use of logic have obtained the best results in their respective subtasks.

Future work aims at developing an Answer Validation model where the hypotheses can include the type of answer requested by the question in order to reformulate the Answer Validation Exercise for the next campaign. Finally, we want to quantify the gain in performance that the Answer Validation systems give in chain with the Question Answering ones.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Technology within the R2D2-SyEMBRA project (TIC-2003-07158-C04-02). 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) and to the people that built the patterns for the hypotheses: Juan Feu (Dutch), Petya Osenova (Bulgarian), Christelle Ayache (French), Bodgan Sacaleanu (German) and Diana Santos (Portuguese).

References

1. R. Bar-Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, I. Szpektor. The Second PASCAL Recognising Textual Entailment Challenge. *In Proceedings of the Challenges Workshop*, pages 1-9, Venice, April 2006.
2. I. Dagan, O. Glickman, and B. Magnini. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising TextualEntailment, Southampton, UK*, pages 1–8, April 2005.
3. J. Herrera, A. Peñas, and F. Verdejo. Question Answering Pilot Task at CLEF 2004. *In Multilingual Information Access for Text, Speech and Images. CLEF 2004*, Volume 3491 of LectureNotes in Computer Science, pages 581–590, 2005.
4. B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo, and M. de Rijke. The Multiple Language Question Answering Track at CLEF 2003. In C. Peters, J. Gonzalo, M. Braschler, and M. Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems. CLEF 2003.*, volume 3237 of *Lecture Notes in Computer Science*, pages 471–486, 2004.
5. B. Magnini, A. Vallin, C. Ayache, G. Erbach, A. Peñas, M. de Rijke, P. Rocha, K. Simov, and R. Sutcliffe. Overview of the CLEF 2004 Multilingual Question Answering Track. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images. CLEF 2003.*, volume 3491 of *Lecture Notes in Computer Science*, pages 371–391, 2004.
6. Anselmo Peñas, Álvaro Rodrigo, Felisa Verdejo. SPARTE, a Test Suite for Recognising Textual Entailment in Spanish. *Lecture Notes in Computer Science 3878*, CiCLing'06, pages 275-286, Springer-Verlag, 2006
7. A. Vallin, B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peñas, M. de Rijke, B. Sacaleanu, D. Santos, and R. Sutcliffe. Overview of the CLEF 2005 Multilingual Question Answering Track. In *Proceedings of CLEF 2005*, 2005.