Overview of the Answer Validation Exercise 2007

Anselmo Peñas, Álvaro Rodrigo, Felisa Verdejo

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

Abstract

The Answer Validation Exercise at the Cross Language Evaluation Forum is aimed at developing systems able to decide whether the answer of a Question Answering system is correct or not. We present here the exercise description, the changes in the evaluation methodology with respect to the first edition, and the results of this second edition (AVE 2007). The changes in the evaluation methodology had two objectives: the first one was to quantify the gain in performance when more sophisticated validation modules are introduced in QA systems. The second objective was to bring systems based on Textual Entailment to the Automatic Hypothesis Generation problem which is not part itself of the Recognising Textual Entailment (RTE) task but a need of the Answer Validation setting. 9 groups have participated with 16 runs in 4 different languages. Compared with the QA systems, the results show an evidence of the potential gain that more sophisticated AV modules introduce in the task of QA.

Keywords

Question Answering, Evaluation, Textual Entailment, Answer Validation

1. Introduction

The first Answer Validation Exercise (AVE 2006) [7] was activated last year in order to promote the development and evaluation of subsystems aimed at validating the correctness of the answers given by 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 text. This automatic Answer Validation is expected to be useful for improving QA systems performance [5]. However, the evaluation methodology in AVE 2006 did not permit to quantify this improvement and thus, the exercise has been modified in AVE 2007.

Figure 1 shows the relationship between the QA main track and the Answer Validation Exercise. The main track provides the questions made by the organization and the responses given by the participant systems once they are judged by humans.

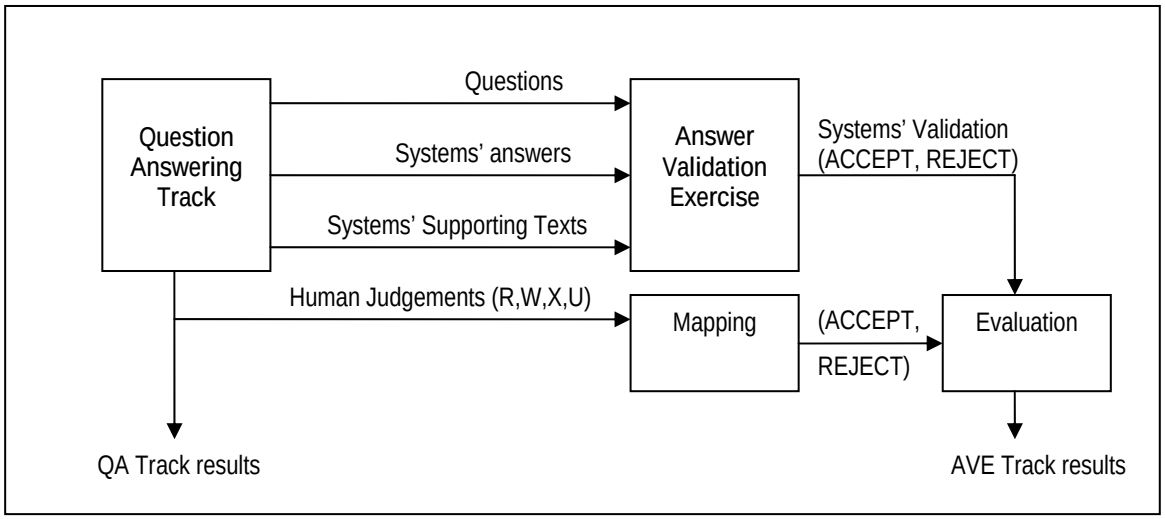

Figure 1. Relationship between the QA Track and the AV Exercise

Another difference in the exercise with respect to the AVE 2006 is the input to the participant systems. Last year we promoted an architecture based on Textual Entailment trying to bring research groups working on machine learning to Question Answering. Thus, we provided the *hypothesis* already built from the questions and answers [6] (see Figure 2). Then, the exercise was similar to the RTE Challenges [1] [2] [3], where systems must decide if there is entailment or not between the supporting text and the hypothesis.

In this edition, on the contrary, we left open the problem of Automatic Hypothesis Generation for those systems based on Textual Entailment. In this way, the task is more realistic and close to the Answer Validation problem, where systems receive a triplet (Question, Answer, Supporting text) instead a pair (Hypothesis, Text) (see Figure 2).

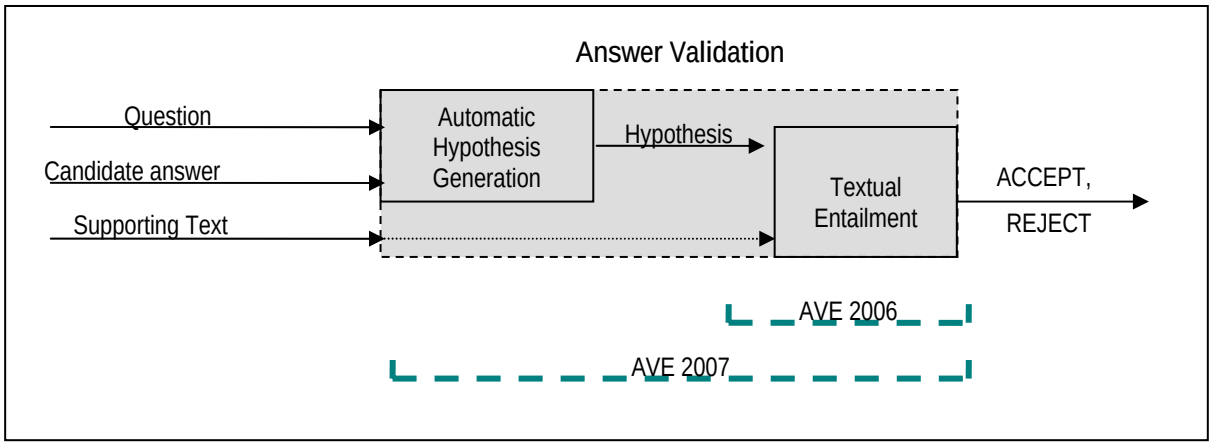

Figure 2. From an Answer Validation architecture based on Textual Entailment in AVE 2006 to the complete Answer Validation systems evaluation in AVE 2007.

Section 2 describes the exercise in more detail. The development and testing collections are described in Section 3. Section 4 discusses the evaluation measures. Section 5 offers the results obtained by the participants and finally Section 6 present some conclusions and future work.

```
<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
                Italian 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 3. Excerpt of the English test collection in AVE 2007

2. Exercise Description

In this edition, participant systems received a set of triplets (Question, Answer, Supporting Text) and they must 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 3). Systems must consider the Question and validate each of the (Answer, Supporting Text) pairs. The number of answers to be validated per question depended on the number of participant systems at the Question Answering main track.

Participant systems must return one of the following values for each answer according to the response format (see Figure 4):

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

Figure 4. Response format in AVE 2007

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

- SELECTED indicates that the answer is VALIDATED and it is the one chosen as the output of 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 no enough evidence of its correctness. There is no restriction in the number of REJECTED answers (from zero to all).

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

3. Collections

Since our objective was to compare AVE results with the QA main track results, we must ensure that we give to AV systems no extra information. The fact of grouping all the answers to the same question could lead to provide extra information based on counting answer redundancies that QA systems might not be considering. For this reason we removed duplicated answers inside the same question group. In fact, if an answer was contained in another answer, the shorter one was removed. Finally, NIL answers, void answers and answers with a supporting snippet larger than 700 characters (maximum permitted in the main track) were discarded for building the collections. This processing lead to a reduction in the number of answers to be validated (see Tables 1 and 2): from 11.2% in the Italian test collection to 88.3% in the Bulgarian development collection.

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

3.1. Development Collections

Development collections were obtained from the QA@CLEF 2006 [6] 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.

	German	English	Spanish	French	Italian	Dutch	Portuguese	Bulgarian
Questions	187	200	200	200	192	198	200	56
Answers (final)	504	1121	1817	1503	476	528	817	70
% over available answers	31.5%	62.28%	53.44%	50.1%	47.6%	44%	40.85%	11.67%
VALIDATED	135	130	265	263	86	100	153	49
REJECTED	369	991	1552	1240	390	428	664	21

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

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

3.2. Test Collections

Test collections were obtained from the QA@CLEF 2007 main track. In this edition, questions were grouped by topic [4]. The first question of a topic was 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.).

For the AVE 2007 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.

The change of the task produced a lower participation in the main track because systems were not tuned on time and this fact, together with the consideration of less number of questions and the elimination of redundancies led to a reduction of the evaluation corpora in AVE 2007.

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	Italian	Dutch	Portuguese	Romanian
Questions	113	67	170	122	103	78	149	100
Answers (final)	282	202	564	187	103	202	367	127
% over available answers	48.62%	60.3%	66.35%	75.4%	88.79%	51.79%	30.58%	52.05%
VALIDATED	67	21	127	[1]	16	31	148	45
REJECTED	197	174	424	[1]	84	165	198	58
UNKNOWN	18	7	13	[1]	3	6	21	24

Table 2. Number of questions and answers in the AVE 2007 test collections

4. Evaluation of the Answer Validation Exercise

In [7] was argued why the AVE evaluation is based on the detection of the correct answers. Instead of using an overall accuracy as the evaluation measure, we proposed the use of precision (1), recall (2) and F-measure (3) (harmonic mean) over answers that must be VALIDATED. In other words, we proposed to quantify systems ability to detect whether there is enough evidence to accept an answer.

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

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

[1] Assessments not available at the this report was submited

$$precision = \frac{|\ predicted\_correctly\_as\_SELECTED\_or\_VALIDATED|}{|\ predicted\_as\_SELECTED\_or\_VALIDATED|} \qquad (1)$$

$$recall = \frac{|\ predicted\_correctly\_as\_SELECTED\_or\_VALIDATED|}{|CORRECT\_answers|} \qquad (2)$$

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

However, 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 the QA architecture. Some recent works [5] have shown how the use of textual entailment can improve the accuracy of QA systems. Our aim was to obtain evidences of this improvement in a comparative and shared evaluation.

For this reason, a new measure (4), very easy to understand, was applied 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.

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

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 in (5). We will refer to this measure also as *percentage of the perfect selection* (*normalized_qa_accuracy x 100*).

$$normalized\_qa\_accuracy = \frac{|\ answers\_SELECTED\_correctly|}{|\ questions\_with\_correct\_answers|} \qquad (5)$$

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 (6).

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

5. Results

Nine groups (2 less than the past edition) have participated in four different languages. 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 5 runs respectively.

Tables 4-7 show the results 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 the QA systems). Together with the systems precision, recall and F-measure, the two baselines values are shown: the results of a system that always accept all answers (validates 100% of the answers), and the results of a hypothetical system that validates the 50% of answers.

	German	English	Spanish	Portuguese	Total
Fernuniversität in Hagen	**2**				**2**
U. Évora				**1**	**1**
U. Iasi		**1**			**1**
DFKI		**2**			**2**
INAOE			**2**		**2**
U. Alicante		**2**			**2**
Text Mess project		**2**			**2**
U. Jaén			**2**		**2**
UNED		**1**	**1**		**2**
Total	**2**	**8**	**5**	**1**	**16**

Table 3. Participants and runs per language in AVE 2007

Group	System	F	Precision	Recall
INAOE	tellez_1	0.53	0.38	0.86
INAOE	tellez_2	0.52	0.41	0.72
UNED	rodrigo	0.47	0.33	0.82
UJA	magc_1	0.37	0.24	0.85
100% VALIDATED		0.37	0.23	1
50% VALIDATED		0.32	0.23	0.5
UJA	magc_2	0.19	0.4	0.13

Table 4. Precision, Recall and F measure over correct answers for Spanish

Group	System	F	Precision	Recall
FUH	iglockner_1	0.72	0.61	0.9
FUH	iglockner_2	0.68	0.54	0.94
100% VALIDATED		0.4	0.25	1
50% VALIDATED		0.34	0.25	0.5

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

Group	System	F	Precision	Recall
DFKI	ltqa_2	0.55	0.44	0.71
DFKI	ltqa_1	0.46	0.37	0.62
U. Alicante	ofe_1	0.39	0.25	0.81
Text-Mess Project	Text-Mess_1	0.36	0.25	0.62
Iasi	adiftene	0.34	0.21	0.81
UNED	rodrigo	0.34	0.22	0.71
Text-Mess Project	Text-Mess_2	0.34	0.25	0.52
U. Alicante	ofe_2	0.29	0.18	0.81
100% VALIDATED		0.19	0.11	1
50% VALIDATED		0.18	0.11	0.5

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

Group	System	F	Precision	Recall
UE	jsaias	0.68	0.91	0.55
100% VALIDATED		0.6	0.43	1
50% VALIDATED		0.46	0.43	0.5

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

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, we wanted to obtain some evidence about the

improvement that more sophisticated AV systems could provide to QA systems. Tables 8-11 show the rankings of systems (merging QA and AV systems) according to the QA accuracy calculated only over the subset of questions considered in AVE 2007. With the exception of Portuguese were there is only one participant group, there are AV systems for each language able to achieve more than 70% of the perfect selection. In German and English, the best AV systems obtained better results than the QA systems, achieving a 93% of the perfect selection in the case of German.

In general, the groups that participated in both QA Main Track and AVE, obtained better results with the AV system than with the QA one. This can be due to two factors: Or they need to extract more and better candidate answers, or they do not use their own AV module to rank them properly in the QA system.

Group	System	System Type	QA accuracy	% of perfect selection
Perfect selection		QA	0.59	100%
Priberam		QA	0.49	83.17%
INAOE	tellez_1	AV	0.45	75.25%
UNED	rodrigo	AV	0.42	70.3%
UJA	magc_1	AV	0.41	68.32%
INAOE		QA	0.38	63.37%
INAOE	tellez_2	AV	0.36	61.39%
Random		AV	0.25	41.45%
MIRA		QA	0.15	25.74%
UPV		QA	0.13	21.78%
UJA	magc_2	AV	0.08	13.86%
TALP		QA	0.07	11.88%

Table 8. Comparing AV systems performance with QA systems in Spanish

Group	System	System Type	QA accuracy	% of perfect selection
Perfect selection		QA	0.54	100%
FUH	iglockner_2	AV	0.50	93.44%
FUH	iglockner_1	AV	0.48	88.52%
DFKI	dfki071dede	QA	0.35	65.57%
FUH	fuha071dede	QA	0.32	59.02%
Random		AV	0.28	51.91%
DFKI	dfki071ende	QA	0.25	45.9%
FUH	fuha072dede	QA	0.21	39.34%
DFKI	dfki071ptde	QA	0.05	9.84%

Table 9. Comparing AV systems performance with QA systems in German

Group	System	System Type	QA accuracy	% of perfect selection
Perfect selection		QA	0.3	100%
DFKI	Itqa_2	AV	0.21	70%
Iasi	adiftene	AV	0.21	70%
UA	ofe_2	AV	0.19	65%
U.Indonesia	CSUI_INEN	QA	0.18	60%
UA	ofe_1	AV	0.18	60%
DFKI	Itqa_1	AV	0.16	55%
UNED	rodrigo	AV	0.16	55%
Text-Mess Project	Text-Mess_1	AV	0.15	50%
DFKI	DFKI_DEEN	QA	0.13	45%
Text-Mess Project	Text-Mess_2	AV	0.12	40%
Random		AV	0.1	35%
DFKI	DFKI_ESEN	QA	0.04	15%
Macquarie	MQAF_NLEN_1	QA	0	0%
Macquarie	MQAF_NLEN_2	QA	0	0%

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

Group	System	System Type	QA accuracy	% of perfect selection
Perfect selection		QA	0.74	100%
Priberam		QA	0.61	82.73%
UE	jsaias	AV	0.44	60%
Random		AV	0.44	60%
U. Evora	diue	QA	0.41	55.45%
LCC	lcc_ENPT	QA	0.3	40%
U. Porto	feup	QA	0.23	30.91%
INESC-ID	CLEF07-2_PT	QA	0.13	17.27%
INESC-ID	CLEF07_PT	QA	0.11	15.45%
SINTEF	esfi_1	QA	0.07	10%
SINTEF	esfi_2	QA	0.04	5.45%

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

All the participant groups in AVE 2007 reported the use of an approach based on Textual Entailment. 5 of the 9 groups (FUH, U. Iasi, INAOE, FUH, U. Évora and DFKI) have also participated in the Question Answering Track, showing that techniques developed for Textual Entailment are in the process of being incorporated in the QA systems participating at CLEF.

	adiftene	**tellez**	**rodrigo**	**iglockner_1**	**iglockner_2**	**jsaias**	**ltqa**	**magc**	**ofe**	**text_mess**
Generates hypotheses	X	X		X	X				X	X
Wordnet	X			X	X					
Chunking		X				X		X		
n-grams, longest common Subsequences		X					X	X	X	X
Phrase transformations	X	X								
NER	X	X	X					X		X
Num. expressions	X	X	X		X	X				X
Temp. expressions			X		X	X				X
Coreference resolution				X	X					
Dependency analysis	X						X		X	
Syntactic similarity	X	X					X		X	
Functions (sub, obj, etc)	X					X	X			
Syntactic transformations	X									
Word-sense disambiguation				X	X					
Semantic parsing	X			X	X	X				
Semantic role labeling				X	X					
First order logic representation				X	X	X				
Theorem prover				X	X	X				
Semantic similarity	X					X				

Table 12. Techniques, resources and methods used by the AVE participants.

Table 12 shows the techniques used by AVE participant systems. In general, the groups that performed some kind of syntactic or semantic analysis worked in the Automatic Hypothesis Generation as a combination of the question and the answer. However, in some cases the hypothesis generated was directly in a logic form instead of a textual sentence.

All the participants reported the use of lexical processing. Lemmatization and part of speech tagging were commonly used. In the other side, only few systems used first order logic representations, performed semantic analysis and took the validation decision with a theorem prover.

Lexical similarity was the feature most used for taking the validation decision. In general, systems that performed syntactic or semantic processing used this processing as similarity features. None of the systems reported the use of semantic frames.

6. Conclusions

In this second edition of the Answer Validation Exercise, techniques developed for Recognizing Textual Entailment have been employed widely, although the exercise was defined more closely to the real answer validation application.

We have refined the evaluation methodology in order to consider the QA systems performance as a reference for AV systems evaluation. Thus, new measures have been defined together with their respective baselines: *qa_accuracy* and the *percentage of the perfect selection* (*normalized_qa_accuracy*).

With respect to the development of test collections, the new evaluation framework led us to reduce redundancies in the sets of answers. This process reduces the size of the testing collections discarding around 50% of candidate answers. The training and testing collections resulting from AVE 2006 and 2007 are available at http://nlp.uned.es/QA/ave for researchers registered at CLEF.

Results show that AV systems are able to detect correct answers improving the results of QA systems. In fact, except for Portuguese (where there is only one participant at AVE), all the systems are far from the random behaviour and closer to the perfect selection (from 70% to 93%).

All systems utilize lexical processing, most of them introduce a syntactic level and only few make use of semantics and logic. Groups that participated in both QA and AVE tracks show better performance in the selection of answers than the results obtained by the whole QA system. This fact points to the need of considering the evidences given by the AV modules in order to generate more and better candidate answers. In this way, the approach of looping the AV module with the generation of candidate answers should be considered instead of the solely approach based on the ranking of candidate answers.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Technology within the Text-Mess-INES project (TIN2006-15265-C06-02), 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. Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini and Idan Szpektor. 2006. The Second PASCAL Recognising Textual Entailment Challenge. In Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venice, Italy.
2. Ido Dagan, Oren Glickman and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. Lecture Notes in Computer Science, Volume 3944, Jan 2006, Pages 177 - 190.
3. Danilo Giampiccolo, Bernardo Magnini, Ido Dagan and Bill Dolan. The Third PASCAL Recognizing Textual Entailment Challenge. ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. 2007.
4. Danilo Giampiccolo et al. 2007. Overview of the CLEF 2007 Multilingual Question Answering Track. Working Notes of CLEF 2007.
5. 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
6. Bernardo Magnini, Danilo Giampiccolo, Pamela Forner, Christelle Ayache, Valentin Jijkoun, Petya Osenova, Anselmo Peñas, Paulo Rocha, Bogdan Sacaleanu, and Richard Sutcliffe, 2007. Overview of the CLEF 2006 Multilingual Question Answering Track. CLEF 2006, Lecture Notes in Computer Science LNCS 4730. Springer-Verlag, Berlín
7. Anselmo Peñas, Álvaro Rodrigo, Valentín Sama, Felisa Verdejo, 2007. Overview of the Answer Validation Exercise 2006. CLEF 2006, Lecture Notes in Computer Science LNCS 4730. Springer-Verlag, Berlín