Testing the Reasoning for Question Answering Validation

Anselmo Peñas, Álvaro Rodrigo, Valentín Sama, Felisa Verdejo

Depto. Lenguajes y Sistemas Informáticos, UNED
Juan del Rosal, 16; 28040 Madrid; Spain
{anselmo,alvarory,vsama,felisa}@lsi.uned.es
Tel. +34 91 398 7750

Abstract

Question Answering (QA) is a task that deserves more collaboration between Natural Language Processing (NLP) and Knowledge Representation (KR) communities, not only to introduce reasoning when looking for answers or making use of answer type taxonomies and encyclopedic knowledge, but also, as discussed here, for Answer Validation (AV), that is to say, to decide whether the responses of a QA system are correct or not. This was one of the motivations for the first Answer Validation Exercise at CLEF 2006 (AVE 2006). The starting point for the AVE 2006 was the reformulation of the Answer Validation as a Recognizing Textual Entailment (RTE) problem, under the assumption that a hypothesis can be automatically generated instantiating a hypothesis pattern with a QA system answer. The test collections that we developed in seven different languages at AVE 2006 are specially oriented to the development and evaluation of Answer Validation systems. We show in this article the methodology followed for developing these collections taking advantage of the human assessments already made in the evaluation of QA systems. We also propose an evaluation framework for AV linked to a QA evaluation track. We quantify and discuss the source of errors introduced by the reformulation of the Answer Validation problem in terms of Textual Entailment (around 2%, in the range of inter-annotator disagreement). We also show the evaluation results of the first Answer Validation Exercise at CLEF 2006 where 11 groups have participated with 38 runs in 7 different languages. The most extensively used techniques were Machine Learning and overlapping measures, but systems with broader knowledge resources and richer representation formalisms obtained the best results.

Keywords: Textual Entailment; Test Collections; Question Answering; Answer Validation; Evaluation

1. Introduction

Behind the Natural Language interface of a Question Answering (QA) system, a very complex processing is performed involving question and answer type classification, text retrieval, answer extraction, answer validation, etc. This complex processing requires not only some kind of text treatment but also involves the use of large resources such as wordnets, gazetteers, paraphrase collections, etc. The acquisition, the management and the representation of this knowledge permitting effective inferences in order to be used by a QA system are still open challenges. For this reason, the work on semantics useful for QA would be enriched by the collaboration between Knowledge Representation (KR) and Natural Language Processing (NLP) communities. In fact, the best performing systems have already incorporated some kind of inference or reasoning in order to identify the candidate answers to a question [7] [13].

At the same time, the experiences at the QA Track of the Cross-Language Evaluation Forum[1] (CLEF) show a lack in the task of Answer Validation. Taking the evaluation of the QA in Spanish during 2005 as an example, 73% of the questions could be answered by at least one participant (see combination column in *Figure 1*), but the best performing system only answered correctly 42% of them [17]. Why is the difference so big? Of course, different systems implement different techniques, and systems that perform better on one type of questions perform worse on others (see *Figure 1*). However, the fact is that we are unable to combine systems results in order to achieve the best potential accuracy. QA systems need further criteria to decide whether their answers are correct or not and, even more important, to assess an accurate confidence score to them. Here is where the need of Answer Validation technologies arises and reasoning is needed. QA systems should be able to assess whether they are providing right answers or not.

An Answer Validation system receives the triplet *Question, Candidate Answer and Supporting Text* and returns a Boolean value indicating if the Answer is correct for the *Question* according to the *Supporting Text* or not (see *Figure 2*). The first Answer Validation Exercise (AVE 2006) was launched at CLEF to promote the development and evaluation of these subsystems aimed at validating the correctness of the answers given by QA systems. Here we describe the evaluation framework, the test collections and the results for AVE 2006.

The *validation of answers* is a task that introduces another chance for reasoning and, therefore, collaboration among KR and NLP communities: the text that supports the answer must *entail* somehow the given answer. In this way, the idea of Answer Validation can be reformulated as a problem of Textual Entailment (RTE) [1] [4] after the step of generating a hypothesis (see *Figure 2*). Thus, AVE 2006 was also useful for the development and evaluation of RTE systems.

[1] http://www.clef-campaign.org

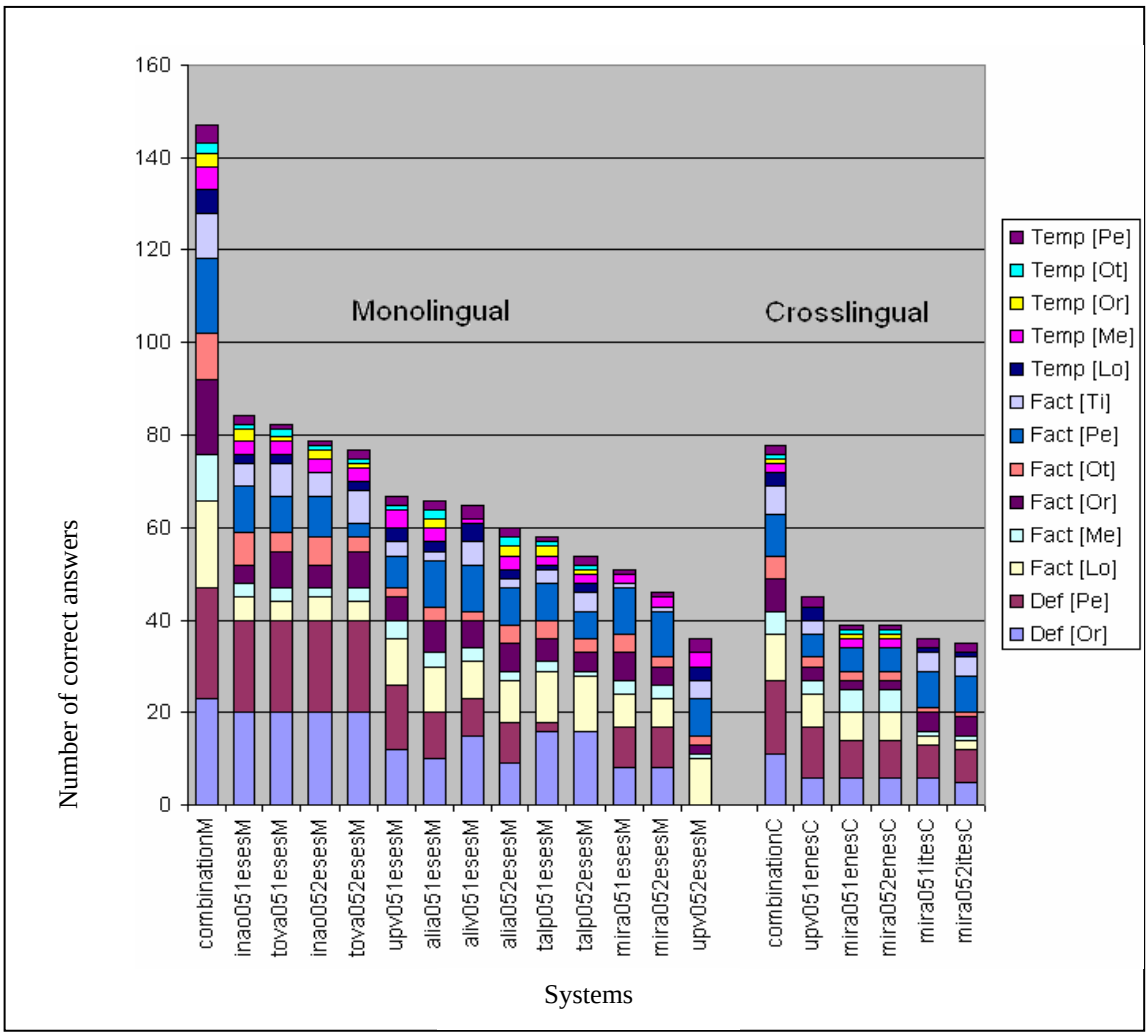

Figure 1. Correct answers in Spanish QA Task at CLEF 2005. Breakdown by type of question: Def – definition, Fact – factoid, Temp – factoid with temporal restriction, Pe – person, Or – organization, Me – measure, Lo – location, Ti – time, Ot – Other.

In *Section 2*, Answer Validation is formulated in terms of Textual Entailment. Then, *Section 3* describes the context of the Answer Validation Exercise. The collections we have developed and the methodology followed for that are described in *Section 4*. *Section 5* explain the evaluation measures we adopted at AVE. The results obtained by different systems in different languages are shown in *Section 6*. In *Section 7* we discuss the fundamental assumptions stated in AVE 2006 and some lessons learned. Finally *Section 8* offers some conclusions and ideas for future work.

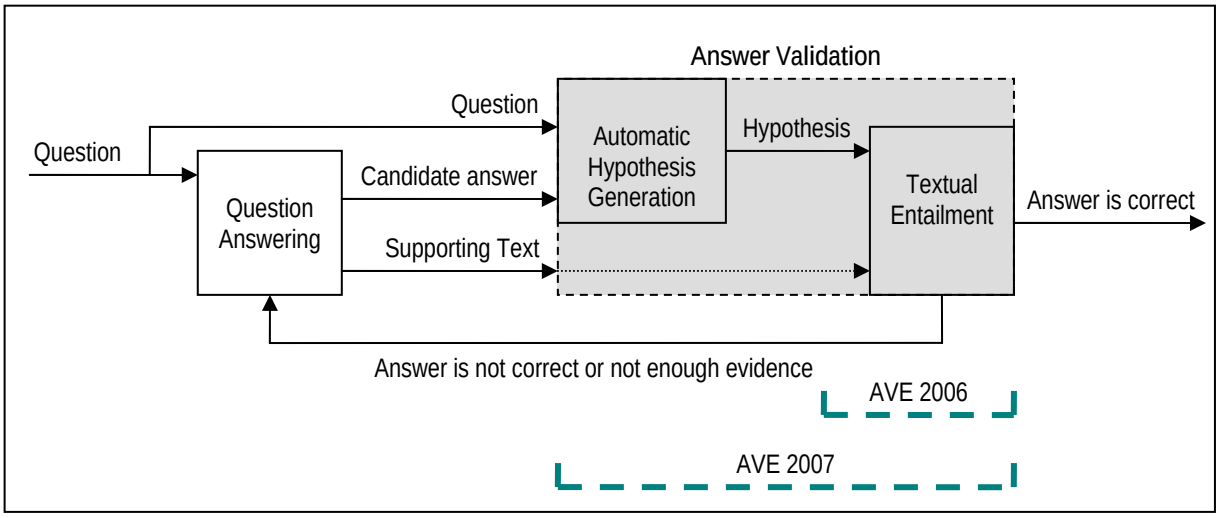

Figure 2. Context, architecture and decision flow for an Answer Validation system based on Recognizing Textual Entailment. AVE 2006 evaluated RTE subsystems while AVE 2007 will evaluate complete Answer Validation systems.

2. Answer Validation as a Textual Entailment problem

The task of Recognizing Textual Entailment (RTE) [4] aims at deciding whether the truth of a text entails the truth of another text named hypothesis or, in other words, if the meaning of the hypothesis is enclosed in the meaning of the text. The entailment relation between texts is useful for a variety of tasks as, for example, Automatic Summarization, where a system could eliminate the passages whose meaning is already entailed by other passages; or Question Answering (QA), where the answer of a question must be entailed by the text that supports the correctness of this answer. This application of RTE to QA can enhance the accuracy of open-domain QA systems [6].

The evaluation of RTE systems has been addressed in the PASCAL RTE Challenges of 2005 and 2006 [1] [4]. In these challenges, the participant RTE systems received a set of text-hypothesis pairs as input and they had to decide whether the text entailed the hypothesis or not, returning YES or NOT as output for each pair. *Figure 3* shows some examples from the second RTE Challenge (annotated test corpus).

The PASCAL RTE data was derived from several different application oriented data [4], by suitable procedures depending on the source task of the pairs. Following [4], when QA systems return both an answer and a text that supports the correctness of the answer, then an RTE system can be applied to validate the answers after this reformulation:

1. Build a hypothesis turning the question plus the answer into an affirmative form. For example, from the question “Who is Vicente Fox?”, and the answer “President of Mexico”, a possible hypothesis could be “Vicente Fox is the President of Mexico”.
2. Evaluate the entailment: If the supporting text entails this hypothesis, then the answer is expected to be correct (see examples of *Figure 3*).

AVE 2006 extended this approach to seven languages, focusing the evaluation on the output of real QA systems.

```
<pair id="286" entailment="YES" task="QA">
       <t>
       President Vicente Fox heads into his final year of office
       Thursday, promising a more democratic, less corrupt and
       economically stable Mexico.
       </t>
       <h>Vicente Fox is the President of Mexico.</h>
</pair>
<pair id="296" entailment="NO" task="QA">
       <t>
       FTAA is a huge extension of the principles behind NAFTA,
       aimed at increasing international government, not
       strengthening the United States.
       </t>
       <h>NAFTA is a huge extension of FTAA.</h>
</pair>
```

Figure 3. Examples of text-hypothesis pairs from the Second PASCAL Recognizing Textual Entailment Challenge.

3. Answer Validation Exercise

The Answer Validation Exercise (AVE) is linked to the Question Answering track as it is shown in *Figure 4*. The test collections are derived semi-automatically taking advantage of the QA systems’ responses and their human assessments.

AVE aims at the evaluation of complete Answer Validation systems that take as input the triplet [*Question, Answer, Supporting Text]* and return a *[YES | NO]* value indicating whether the *Answer* to the *Question* is correct or not according to the *Supporting Text* (see *Figure 2*). In this way, systems based on Textual Entailment must take into account the problem of Automatic Hypothesis Generation which is a new problem in the context of QA. From a theoretical point of view, the main difficulty is to ensure that a generated hypothesis will confirm that the answer is correct if this hypothesis is validated.

In the first AVE (2006), we decided to exclude the Automatic Hypothesis Generation problem and provide hypotheses already constructed to the participants, in a similar way to that proposed by the PASCAL RTE Challenges.

A possible approach for the Automatic Hypothesis Generation is to obtain a pattern (related to the question) that instantiated with an answer generates the hypothesis. The experience of AVE 2006 permitted us to:

1. Test that this approach is feasible, although there are some limitations that we discuss in Section 7.
2. Generate 200 pairs (question, hypothesis-pattern) for each language that can be useful for training the automatic hypothesis generation subsystems.

In this way we leave prepared the evaluation of complete Answer Validation systems in AVE 2007.

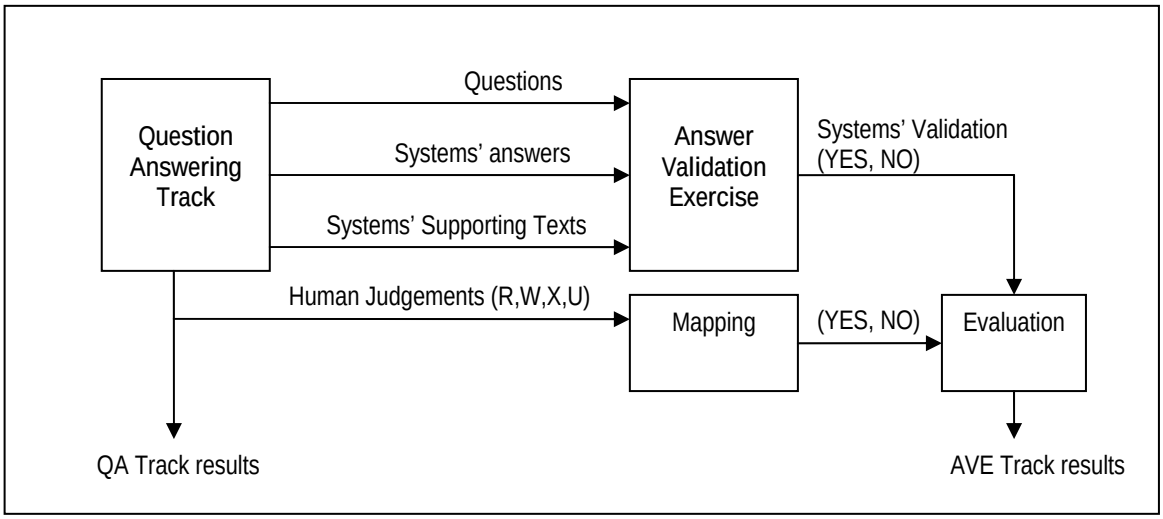

Figure 4. Relationship between the QA Track and the AV Exercise

4. Collections

Since RTE task has been defined recently, there exist only few corpora for training and testing RTE systems, all of them in English and not completely focused on Question Answering. The AVE 2006 collections are inspired in the corpus used for training and testing RTE systems at the PASCAL RTE Challenges [1] [4]. These corpora are available[2] only in English. In the PASCAL RTE collections the number of pairs with entailment (TRUE/YES) is equal to the number of pairs without entailment (FALSE/NO). All these pairs were selected, filtered and adapted manually from a different number of sources related to different areas such as Information Retrieval, Machine Translation, Information Extraction, Paraphrase Acquisition or Question Answering among others. All the text-hypothesis pairs have a tag indicating the corresponding type of source. In the first PASCAL RTE Challenge, MITRE decided to cast the RTE problem as one of statistical alignment and for this, they needed a corpus larger than the one provided by the organization [3]. From the MITRE point of view, most of the TRUE

[2] Available at http://www.pascal-network.org/Challenges/RTE/

pairs exhibit a paraphrase relationship in which the hypothesis is a paraphrase of a subset of the text. Therefore, they took a news corpus in which the headline of a news article is often a partial paraphrase of the lead paragraph. After a semi-automatic processing they selected the most promising 100,000 pairs, estimating that 74% of them have an entailment relationship. This corpus lead MITRE to one of the best results in the first PASCAL Challenge. Although this is a corpus useful for training statistical RTE systems, the absence of human assessment for every pair do not permit its use for evaluation purposes. There are some other works aimed at acquiring paraphrases without the notion of entailment [2] [10] [16]. One corpus available in this direction is the Microsoft Research Paraphrase Corpus[3] [5]. Based on the idea that an event generates hundreds of different news articles in a closed period of time, they decided to apply unsupervised techniques over news clusters for acquiring sentence-level paraphrases. Again, this corpus could be useful for training RTE systems working with English rather than for evaluation purposes.

The AVE development collections reuse the questions, answers and human assessments of the CLEF QA Track during 2003, 2004 and 2005 [8] [11] [12] [15] [17]. Two different collections were developed for the two languages with higher number of participants at QA: Spanish and English. The AVE test collections were produced from the questions and the systems' responses in seven different languages at CLEF QA 2006.

The main difference between the development and the test collections in AVE 2006 comes from the availability of the supporting text snippets. In the previous editions of CLEF, systems did not provide the supporting text snippet but only the document identification, and therefore, we had to extract them automatically. Since the QA human assessments did not consider these snippets, this step introduced errors in the development collections that were quantified and corrected.

In CLEF 2006 systems had to return the supporting text together with the answer, and therefore they were considered in the human assessments. Thus, no errors were introduced in the AVE test collections by an artificial identification of the supporting texts.

The following subsections describe the methodology we followed for building the AVE collections.

4.1. Building the hypothesis

Since Textual Entailment is defined between statements, the first step was to turn the questions into an affirmative form. For example, question (1) was transformed manually into pattern (2)

(1) "Which is the capital of Croatia?"

(2) "<answer/> is the capital of Croatia"

where the mark "<answer/>" has to be instantiated with any answer given to that question by any system. From these patterns we built all the hypotheses automatically by substituting the mark "*<answer/>*" with the

[3] Available at http://research.microsoft.com/research/downloads/default.aspx

corresponding answers given by the real QA systems. For example, for the answer *"Zagreb"* the instantiation of the pattern would give the hypothesis *"Zagreb is the capital of Croatia"*.

Some wrong or inexact answers could provide the hypothesis not only a wrong semantics but also a wrong syntactic structure. For example, for the answer *"Zagreb was then seen as the political center"* the instantiation of the pattern would give the hypothesis *"Zagreb was then seen as the political center is the capital of Croatia"*. Several pairs without entailment in the collections have this feature. Since the final objective is the development of real Answer Validation subsystems, this is a desirable feature for the corpus, allowing the development of syntactic criteria for detecting wrong answers, and also promoting the development of systems robust enough to some formal and syntactic errors.

Finally, repeated answers (instances) and NIL answers were removed. NIL answers might be correct or not, but in any case, NIL stands for the absence of answer and therefore, there is no answer to validate.

4.2. Identifying the text snippet

In 2003, 2004 and 2005 CLEF campaigns no supporting text snippets were requested for the answers, but only the identification of the whole document. Since the formulation of the validation as a RTE problem requires these snippets, we had to identify them automatically and then revise them manually for the development collections. The processing was very naive but it identified the correct snippet in 81% of the cases:

1. The question was tokenized eliminating stop words.
2. The document indicated in the answer was segmented by sentences.
3. The sentence containing the answer string and more question words (more overlapping) was chosen.
4. If no sentence in the document contains the answer string, then the sentence with more words in common is chosen.
5. The following sentences up to 500 characters (maximum size allowed for the text snippets) were included in the final excerpt.

4.3. Determining the entailment value

The QA assessments were not binary values and, therefore, a simple mapping was necessary for converting QA assessments into entailment values:

- *Correct answer (R):* the answer to the question is correct and the document supports its correctness. Then the text entails the hypothesis and the entailment value is YES, conditioned to the automatically extracted snippet.

- *Unsupported answer (U):* although the answer could be considered as correct, the document does not allow affirmation of the correctness of the answer. Thus, no text snippet of the document can entail the hypothesis, and therefore the entailment value is NO.
- *Inexact answer (X):* This is a difficult case also for the human assessors of Question Answering. There is no additional information to decide whether the answer contains too much information or, in the contrary, the answer string is too short to be considered as responsive. Both cases were tagged as *Inexact*. Thus, we do not have clear criteria to determine the entailment values in these cases without a human assessment. Fortunately, there are only few cases (6% of the pairs) and we opted for ignoring them in the evaluation.
- *Incorrect answer (W):* the answer to the question is wrong. Although the answer could be directly extracted from the text, the joint reformulation of question and answer as a statement (hypothesis) avoids the entailment between the text and the hypothesis. In fact, we evaluated that only 2% of the wrong answers could produce text-hypothesis with entailment (see *Tables 2 and 3*, and *Section 7* for the discussion). Therefore, we marked the entailment value in these cases as NO entailment.

4.4. Development collections

Table 1 shows the number of question and answer pairs available at the beginning of the Spanish collection development. Each answer was assessed by humans during the CLEF campaigns in order to decide whether it was correct, exact and supported by the given document or not.

Year	**#questions**	**#answers**	**#runs**	**#participants**	**#q-a pairs**
2003	200	3	2	1	1200
2004	200	1	2	5	1600
pilot 2004	100	1	1	1	100
2005	200	1	2	9	3598
				Sum	6498

Table 1. Number of question answer pairs used for the Spanish development collection

There exist several sources of potential errors that we wanted to quantify in order to avoid them in the construction of the test collections. We performed a partial human evaluation of the Spanish development corpus in order to assess its quality. Table 2, shows the errors found in the development collections for the Spanish exercise before it was delivered.

	Revised	Errors in QA assessments	Reformulation from QA to Text-Hypothesis	Errors in automatic snippet extraction	Total
YES pairs	100%	6%	2%	13%	21%
NO pairs	5%	4%	2%	-	6%

Table 2. Errors found in the development collections for the Spanish AVE

Since the AVE 2006 evaluation measures were aimed at quantifying the detection of the pairs with entailment (pairs YES) and only them (see Section 4.3) we revised 100% of the pairs tagged with entailment value YES (695 pairs), 100% of the pairs tagged as UNKNOWN (65 pairs) due to *Inexact* or not assessed answers, and only 5% of the pairs tagged as NO (113 pairs). The result of the evaluation was the following:

- *Errors in the original QA assessments.* In general, these errors must be in the range of the inter-annotator disagreement. However, we found in the development collection 6% of errors in pairs YES, which was higher than we expected. These errors were due mainly to answers not supported by the document but assessed erroneously as *Right.* In the testing collections, the assessment was easier since judges only had to take into account the given snippet. Thus, in the test collection errors come down to the inter-annotator disagreement ranges (2%).
- *Errors in the automatic identification of the text snippets.* We found that in 100% of the cases one sentence of the text was enough to support the answer or, in other words, to entail the hypothesis. However, the automatic processing failed in 13% of the pairs, giving a snippet that did not support the answer. These pairs were transformed into NO pairs before the releasing of the collection. These errors did not affect the test collections since the human assessment in 2006 considered the given snippets.
- *Errors due to the reformulation of Answer Validation in terms of Recognizing Textual Entailment*. There is an error around 2% due to wrong answers that can produce pairs with entailment. Although the rate is low, these kinds of errors deserve more attention because they affect the definition of the task (see *Section 7*).
- *Re-assessment of Inexact answers*. Assessors in QA always have problems with the *Inexact* answers, where the decision becomes more subjective. However, once the answer was formulated as a hypothesis, it was easier for the assessors to decide if there was entailment or not. In the development collection 32% of the UNKNOWN pairs became YES pairs and 68% became NO pairs, whereas in the test collection 42% of the UNKNOWN pairs became YES and 52% became NO.

4.5. Test Collections

As a difference with the previous campaigns, in 2006 a text snippet was requested to support the correctness of the answers in the QA Track. Thus, the test collections were free of the errors introduced by the automatic identification of the snippets. The QA assessments were done considering the given snippet, so the direct relation between QA assessments and RTE judges was preserved. Table 3 shows the errors found in the Spanish test collection after revising the 100% of pairs.

	Revised	Errors in QA assessments	Reformulation from QA to Text-Hypothesis	Errors in automatic snippet extraction	Total
YES pairs	100%	2.5%	1%	-	3.5%
NO pairs	100%	2.5%	0.4%	-	2.9%

Table 3. Errors found in the test collection for the Spanish AVE

Table 4 shows the number of pairs for each language obtained as the result of the methodology described. Pairs tagged with an entailment value equal to UNKNOWN come from answers judged as *Inexact* or from answers not evaluated at the QA Track. These pairs are ignored in the systems' performance evaluation.

	German	**English**	**Spanish**	**French**	**Italian**	**Dutch**	**Portuguese**
YES pairs	353(25%)	215(10%)	671(28%)	705(22%)	187(16%)	81(10%)	188(14%)
NO pairs	1053(73%)	1144(55%)	1615(68%)	2359(72%)	901(79%)	696(86%)	604(46%)
UNKNOWN	37(2%)	729(35%)	83(4%)	202(6%)	52(5%)	30(4%)	532(40%)
Total	1443	2088	2369	3266	1140	807	1324

Table 4. YES, NO and UNKNOWN pairs in the test collections of AVE 2006

Percentages of YES, NO and UNKNOWN pairs are similar in all languages except for the percentage of UNKNOWN pairs in English and Portuguese. In these languages, up to 5 runs in the QA task were not assessed and therefore, the corresponding pairs could not be used to evaluate the AVE systems.

Development and testing collections resulting from the first AVE 2006 are available at http://nlp.uned.es/QA/ave for researchers registered at CLEF.

5. Evaluation measures

The evaluation at AVE 2006 was focused on the detection of correct answers and only them or, in other words, on the detection of the pairs with entailment. There are two reasons for this approach:

1. 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 the 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. We think this leads to different system development strategies.
2. 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. Although numbers are expected to evolve, the important point is that the evaluation of Answer Validation modules must consider the real output of Question Answering systems, which is not balanced. If we 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.

Therefore, instead of using an overall accuracy as the evaluation measure, we propose 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 propose to quantify the systems' ability to detect the pairs with entailment or to detect whether there is enough evidence to accept an answer.

$$precision = \frac{|\ predicted\_as\_YES\_correcly\ |}{|\{predicted\_as\_YES\}|} \qquad (1)$$

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

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

On the other hand, the higher the proportion of YES pairs is, the higher the baselines are. Thus, results can be compared between systems but considering the baseline of returning always a YES value.

Although UNKNOWN pairs were present in the test collection they were ignored in the evaluation. The cases where systems assessed a YES value to the UNKNOWN pairs did not count for the precision calculation.

6. Results in AVE 2006

Eleven groups have participated in the AVE 2006 in seven different languages. Table 5 shows the participant groups and the number of runs they submitted per language. System reports can be found in the CLEF 2006 Working Notes [14] (the main author of each report is written in parenthesis when the report is available). At least two different groups participated in each language, so the comparison between different approaches is possible. English and Spanish were the most popular with 11 and 9 runs respectively.

	German	English	Spanish	French	Italian	Dutch	Portuguese	Total
Fernuniversität in Hagen (FUH) (Glöckner)	2							2
Language Computer Corporation (LCC) (Tatu)		1	1					2
U. Rome "Tor Vergata" (Zanzotto)		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 (Bosma)	1	2	2	1	1	2	1	10
UNED (Herrera)			2					2
UNED (Rodrigo)			1					1
ITC-irst (Kouylekov)		1						1
R2D2 project			1					1
Total	5	11	9	4	3	4	2	38

Table 5. Participants and runs per language in AVE 2006

Only 3 out of the 12 developer groups (FUH, LCC and ITC-IRST) have participated in the Question Answering Track, showing the chance of collaboration between different communities in benefit of the QA technologies development. We expect that in the near future the QA systems will take advantage of these communities working on the kind of reasoning needed for the Answer Validation.

Tables 6 show the results for all participant systems in each language together with the techniques reported by each group. The results of a system that always accepts all answers (returns YES in 100% of the pairs) is given as a baseline. We also provide the results of a hypothetical system that returns YES at random for 50% of pairs.

The values in the table correspond to the best F-measure obtained per system and language, and (in parenthesis) the gain with respect to the baseline. Since the number of pairs and the proportion of the YES pairs are different for each language, results can not be compared between languages.

The most extensively used techniques were Machine Learning and overlapping measures between text and hypothesis. Systems that reported the use of Logic showed a very good performance. At least one of them (COGEX from LCC) utilizes large amounts of knowledge that, according to the conclusions in RTE-2, seems to be a critical factor for system success.

	German	**English**	**Spanish**	**French**	**Italian**	**Dutch**	**Portuguese**	**Methods reported**
Language Computer Corporation (Tatu)		**0.46** (+70%)	**0.61** (+35%)					Logic, Knowledge resources
Fernuniversität in Hagen (Glöckner)	**0.54** (+38%)							Corpus, Syntax, Logic, Lexical, Semantics,
UNED (Herrera)			**0.57** (+27%)					Overlap, ML
UNED (Rodrigo)			**0.53** (+18%)					NE recognition
U. Rome "Tor Vergata" (Zanzotto)		**0.41** (+52%)						ML, Syntax
ITC-irst (Kouylekov)		**0.39** (+44%)						Syntax, ML, Lexical, Corpus
U. Alicante (Kozareva)	**0.47** (+21%)	**0.37** (+37%)	**0.53** (+18%)	**0.47** (+27%)	**0.41** (+41%)	**0.30** (+58%)	**0.15** (-61%)	Overlap, Corpus, ML
R2D2 project			**0.49** (+9%)					Voting, Overlap, ML
U. Alicante (Ferrández)		**0.32** (+19%)						Lexical, Syntax, Logic
LIMSI-CNRS				**0.11** (-70%)				Paraphrase, Lexical, Syntax
U. Twente (Bosma)	**0.14** (-64%)	**0.30** (+11%)	**0.47** (+4%)	**0.09** (-76%)	**0.17** (-41%)	**0.39** (+105%)	**0.35** (-8%)	Syntax, ML, Overlap
U. Politecnica de Valencia		**0.08** (-70%)						ML
100% YES Baseline	**0.39**	**0.27**	**0.45**	**0.37**	**0.29**	**0.19**	**0.38**	-
Random	**0.33**	**0.24**	**0.37**	**0.32**	**0.26**	**0.17**	**0.32**	-

Table 6. AVE 2006 Results. Values correspond to the best F-score obtained per system and language. Gain with respect to the baseline is given in parenthesis.

7. Discussion

The fundamental assumptions of the AVE 2006 were that (i) hypotheses can be generated instantiating patterns with the system answers, and (ii) Answer Validation can be reformulated as an RTE problem after the step of hypothesis generation. These assumptions have been validated in the AVE 2006, though they have some limitations that are discussed in the following sections.

7.1. Hypothesis generation from patterns and answers.

Only 2% of the hypothesis are sentences that can be found exactly in texts. This is the result of constructing the hypothesis as an affirmative expression of the question, producing a new statement not present in the text that will require some kind of inference from systems in order to detect the entailment. In our opinion this is a good feature of the *pattern* approach in the collection construction (from the evaluation point of view).

The effort required for constructing one pattern for each question is low. However, when there are different possible patterns, some of them can derive in the validation of non-responsive answers. For example, we can generate two patterns for the question "Which is the capital of Croatia?":

1. The capital of Croatia is </answer>
2. <answer/> is the capital of Croatia

The first one generates a more natural expression but permits the validation of non-responsive answers. For example, the answer *"placed in the continental part of Croatia"* would be validated (entailment value YES) according to the supporting text snippet *"The capital of Croatia is placed in the continental part of Croatia and has one million inhabitants"*. For this reason, the second pattern seems more robust.

The fact that patterns are built without knowing the answers introduces some difficulties. For example, different answers to the same question might require different prepositions or articles (a year might require the preposition *in* whereas a day might require the article *the*). Thus, the election of some prepositions in the hypothesis patterns could introduce some noise in the syntax of the hypothesis, although humans are able to recover the appropriate semantics. A solution here would be to request QA systems to answer correct and complete noun or prepositional phrases.

7.2. Pairs tagged as NO that can be considered YES after the RTE reformulation.

We found that some pairs are tagged with entailment value NO while in fact there exists an entailment according to the text. In other words, although the source answers can be considered non responsive, and in fact the corresponding pairs have been tagged with an entailment value NO, the corresponding hypotheses might be true according to the texts. This source of errors is in the same range as the disagreement between the human annotators (2%) that made the QA assessments [17].

Figure 6 shows some examples (translated into English) of this kind of pairs found in the Spanish test collection of AVE 2006. For example, in the third pair, the system answered the name of the *Person* that holds the record instead of the *Measure* of the record.

Question:	What is Deep Blue?
Answer:	developed by IBM
QA assessment:	**Wrong**
Hypothesis:	Deep Blue is developed by IBM
Supporting text:	.. Deep Blue, developed by IBM, was the first machine to win a chess game against a reigning world champion (Garry Kasparov)...
Entailment	**YES**
Question:	Who is the General Secretary of Interpol?
Answer:	General Secretary of the International Criminal Police Organization
QA assessment:	**Wrong**
Hypothesis:	The General Secretary of Interpol is General Secretary of the International Criminal Police Organization
Supporting text:	The computer "X 400" was presented by the General Secretary of the International Criminal Police Organization (INTERPOL) Raymond Kendall...
Entailment	**YES**
Question:	What is the world record in the high jump?
Answer:	obtained by Javier Sotomayor
QA assessment:	**Wrong**
Hypothesis:	The world record in the high jump is obtained by Javier Sotomayor
Supporting text:	... the world record in the high jump, obtained by Javier Sotomayor, is 2.45 metres...
Entailment	**YES**

Figure 6. Non-responsive answers in QA that could be validated via RTE

We also detected that this problem could appear sometimes when the question is related to an event located in time and space. For example, to the question *"Where did the Titanic sink?"* we could generate the pattern *"The Titanic sank in <answer/>"*. Suppose that a system returns as a supported answer "Atlantic Ocean" and other system returns "1912". Both can generate correct hypotheses that are entailed by the text: *"The Titanic sank in the Atlantic Ocean"* and *"The Titanic sank in 1912"*. But only the first one deserves an Answer Validation value of YES.

A possible solution is the consideration of the expected type of answer, creating hypotheses with several clauses that must be entailed. In our example:

The Titanic sank in <answer/> (1)

<answer/> is a location (2)

or, in a compressed way, *"The Titanic sank in <answer type=location/>"*. This gives partial solution to the problem, but requires (from the evaluation perspective):

1. A consensual in the taxonomy of expected answer types which is very difficult since each QA system utilizes its own taxonomy. For example, the taxonomy in [9] is close to the things that a computer can classify but far from the taxonomies acceptable for a documentalist. Again, this kind of work becomes a natural place for collaboration between KR and NLP communities.
2. More sophisticated ways to support the answer. In our example we need a supporting text for the clause *"The Titanic sank in the Atlantic Ocean"*, and another explicit evidence supporting that *"Atlantic Ocean is a location"*. In most of the cases, the text that supports the answer do not contain explicitly this kind of encyclopedic or ontological knowledge of the world, but human QA assessors accept the answer. There is an inference behind regarding ontological knowledge of the world that is ignored by human assessors. This is another place for collaboration. Another examples are the questions asking explicitly *"Which organization did something?"* Once the entity is retrieved, the supporting text does not make explicit that the entity is an organization. For example, *"Which company was acquired by Nokia in 1998?"* could receive a correct and supported answer *"Vienna Systems was acquired by Nokia in 1998"* that should be validated as correct answer, although there is no explicit information about that *"Vienna Systems is a company"*. Here the problem from the evaluation point of view is more difficult. Sometimes it is not easy to decide when the missed piece of information should be required or not in the supporting text.

In conclusion, wide ontologies fully populated and consistent with the answer type taxonomies are needed, deserving more collaboration between our communities.

8. Conclusions and future work

Question Answering is a task that deserves more collaboration between NLP and KR communities, not only to introduce reasoning when looking for answers or making use of answer type taxonomies and encyclopedic knowledge, but also, as discussed here, for Answer Validation. This was one of the motivations for the first Answer Validation Exercise at CLEF 2006 after the experience of the PASCAL RTE Challenges.

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 showed here the methodology for developing the collections taking advantage of the human assessments made in the evaluation of QA systems. We have also proposed a methodology for the evaluation linked to a QA Track.

The AVE 2006 experience permitted us to detect and quantify the source of errors introduced by the reformulation of the problem in terms of Textual Entailment (around 2%, in the range of inter-annotator disagreement).

We also have described the evaluation and results of the first Answer Validation Exercise at CLEF 2006. 11 groups have participated with 38 runs in 7 different languages. Although systems that reported the use of logic have obtained the best results in their respective subtasks, it is not clear whether the advantage was due to the large amount of knowledge that they utilize or due to the fact that knowledge and reasoning were represented in a logic-based framework.

Future work aims at three directions. The first one is to develop a model where the hypotheses can include the expected type of question in a natural and flexible way.

The second direction is to redefine AVE in order to measure the gain in performance that the Answer Validation systems can provide to Question Answering. For this reason, we will move from an intrinsic evaluation in 2006 to an extrinsic evaluation in 2007, in which the best system will be the one that produces more performance gain in Question Answering.

Finally, the third direction will be to include in AVE the chance to study the Automatic Hypothesis Generation problem. In AVE 2007 we will not provide the hypothesis already generated but the original question and the answer string given by the QA systems (see *Figure 2*). Pure RTE systems will be able to participate in the exercise after solving the Automatic Hypothesis Generation problem. At the same time the participation will remain open to other approaches than RTE.

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), the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267) and a PhD grant by UNED. We are grateful to all the people involved in the organization of the QA track (especially to the people 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). Special thanks also to the anonymous reviewers that helped with their comments toward the final editing of the paper.

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. R. Barzilay and L. Lee. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In *HLT-NAACL 2003: Main Proceedings*, pages 16–23, 2003.
3. J. Burger and L. Ferro. Generating an Entailment Corpus from News Headlines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 49–54, June 2005.
4. I. Dagan, O. Glickman, and B. Magnini. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 1–8, April 2005.
5. B. Dolan, C. Quirk, and C. Brockett. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of COLING 2004,Geneva, Switzerland*, 2004.
6. 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
7. S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, J. Williams, J. Bensley: Answer Mining by Combining Extraction Techniques with Abductive Reasoning. *In Proceedings of TREC 2003*, pages 375-382, 2003.
8. 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 Lecture Notes in Computer Science, pages 581–590, 2005.
9. X. Li and D. Roth. Learning Question Classifiers. Proceedings of the 19th International Conference on Computational Linguistics, COLING'02, 2002.
10. D. Lin and P. Pantel. DIRT Discovery of inference rules from text. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328. ACM Press, 2001.
11. 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.
12. 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.
13. Dan I. Moldovan, Christine Clark, Sanda M. Harabagiu, Steven J. Maiorano: COGEX: A Logic Prover for

Question Answering. *In Proceedings of HLT-NAACL 2003*, pages: 87-93, Edmonton, 2003

14. A. Nardi, C. Peters, J.L. Vicedo editors, Working Notes of the CLEF 2006 Workshop, Alicante, Spain, 2006.
15. A. Peñas, F. Verdejo, and J. Herrera. Spanish Question Answering Evaluation. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing. CICLing 2004.*, volume 2945 of *Lecture Notes in Computer Science*, pages 472–483, 2004.
16. Y. Shinyama, S. Sekine, K. Sudo, and R. Grishman. Automatic Paraphrase Acquisition from News Articles. In *Automatic Paraphrase Acquisition from News Articles. Proceedings of Human Language Technology Conference, San Diego, USA.*, 2002.
17. 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.