# UNED at Answer Validation Exercise 2007

Álvaro Rodrigo, Anselmo Peñas, Felisa Verdejo

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

## Abstract

The objective of the Answer Validation Exercise (AVE) 2007 is to develop systems able to decide if the answer to a question is correct or not. Since it is expected that a high percent of the answers, questions and supporting snippets contain named entities, the paper presents a method for validating answers that uses only information about named entities. The promising results aim us to improve the system and use it as a component of other systems.

## Keywords

Named Entities, Question Answering, Textual Entailment, Answer Validation

## 1. Introduction

The Answer Validation Exercise (AVE) 2007 [4] of the Cross Language Evaluation Forum (CLEF) 2007 is aimed at developing systems able to decide whether the responses of a Question Answering (QA) system are correct or not. As a difference with last year [5], the organization does not provide the participants with text-hypothesis pairs in order to decide if there is or not entailment. This year, participant systems receive a set of answers and theirs corresponding supporting snippets grouped by questions. Thus, it is not mandatory the use of textual entailment. Systems must return a value VALIDATED or SELECTED if they considered that the answer is correct and the snippet supports it, and return REJECTED if the answer is incorrect or it is no supported by the text snippet.

The system we have presented is based on the one presented to AVE 2006 [7], which gave good results in Spanish, and the one used in our participation at RTE-3 [6], which obtained also good results in textual entailment over pairs from QA. These two systems were based on named entities (NE). However, these systems needed text-hypothesis pairs that are not given at AVE 2007. This paper shows a system based in named entities that has been adapted to the new specifications at AVE 2007. We have participated with this system in both English and Spanish.

Our main motivation for using named entities is the high percentage of factoids questions in QA@CLEF, representing, for example, 79% of questions in last year Spanish test set [3]. The answers to these questions are expected to be named entities (person names, locations, numbers, dates...) and it is expected that these questions, answers and snippets contain a high amount of named entities.

The main components of the system are described in Section 2. The results and the error analysis are shown in Section 3. Finally, some conclusions and future work are given in Section 4.

## 2. System description

The system receives a set of triplets (question, answer, snippet) and decides, using only information about named entities, if the answer to the question is correct and the text snippet supports it.

As the system uses only information about named entities, the first step is to detect them in a robust way. Then, the second step is the definition and implementation of an entailment relation between named entities.

Next subchapters describe in detail the steps involved in the decision of named entities entailment.

## 2.1 Named entity recognition

Numeric expressions (NUMEX), proper nouns (PN) and time expressions (TIMEX) of questions, answers and snippets are tagged using the FreeLing [1] Name Entity Recognizer (NER). The values of numeric and time expressions are also normalized in order to make easier the entailment decision.

In order to avoid errors in the process of named entities entailment, as it is explained in [7], all named entities receive the same tag NE ignoring the named entity categorization given by the tool.

## 2.2 Named entity entailment

Once the named entities of questions, answers and snippets are detected, the next step is to determine the entailment relations between them.

As it is explained in [6], we consider that a named entity NE1 entails a named entity NE2 if the text string of NE1 contains the text string of NE2. However, some characters change in different expressions of the same named entity as, for example, in a proper noun with different wordings (e.g. Yasser, Yaser, Yasir). To detect the entailment in these situations, when the previous process fails, we implemented a modified entailment decision process taking into account the edit distance of Levenshtein [2]. Thus, if two named entities differ in less than 20%, then we assume that exists an entailment relation between these named entities.

## 2.3 Validation decision

In [6] and [7], we detected the entailment relation between named entities in the text and in the hypothesis. In AVE 2007 [4], this is no possible due to the fact that none hypothesis is given.

As it is described in [5], the hypotheses given by the AVE 2006 organization were build as a combination of questions and answers. This fact aims us to think the possibility of developing a module able to build hypotheses with answers and questions as input. However, as our system needs only the named entities from the hypothesis, we studied how to obtain them without building a textual hypothesis. Our intuition was that the named entities of a certain hypothesis were the same as the named entities of the question plus the named entities of the answer from which the hypothesis is generated.

A look to AVE 2006 corpus shows us that our intuition was correct as figure 1 shows. In the example showed in the figure, the hypothesis has been obtained from the question and answer of the example. The named entities of the hypothesis (Iraq, Kuwait and 1990) correspond to named entities in the question (Iraq, 1990) and the answer (Kuwait).

**Question:** Which country did **<NE>**Iraq**</NE>** invade in **< NE >**1990? **</NE>**
**Answer:** **<NE>**Kuwait**</NE>**

**Hypothesis:** **<NE>**Iraq**</NE>** invaded the country of **<NE>**Kuwait**</NE>** in **<NE>**1990**</NE>**

Figure 1. An example of how the NEs of hypothesis are the NEs of question and answer.

Thus, the validation decision for each triplet (question, answer, snippet) is obtained taking into account the named entities from the text snippet in one way and the named entities from the question plus the named entities from the answer in another way as named entities of a supposed hypothesis.

Then, for taking the final decision, we think that in textual entailment all the elements in the hypothesis must be entailed by elements of the supporting text. Therefore, the system assumes that if there is a named entity in the hypothesis not entailed by one or more named entities in the text, then the answer is not supported or incorrect and then the system must return the value REJECTED for this triplet.

However, in pairs where all the entities in the hypothesis are entailed, there is not enough evidence to decide if the answer is correct or not. In this situation, in order to perform an experiment to obtain some information of the performance of our system, we decided to return the value VALIDATED.

Even though the validation decision describes above shows a good performance in the Spanish development set, the results in English were lower mainly due to errors in the recognition of named entities in the text snippets. An example of these errors is shown in figure 2, where Italy has not been recognized as a named entity in the text snippet.

**Question:** What is the name of the national airline in **<NE>**Italy**</NE>**?
**Snippet:** Italy 's national airline **<NE>**Alitalia**</NE>**

Figure 2. An example of a NE recognition error.

Then, we thought that in our validation decision process it was important that the named entities of the hypothesis (combination of question and answer) were entailed by elements in the text snippet, without the necessity that these elements were recognised as named entities. In order to study this approach, an experiment was performed over the English development set with two different systems:

1. A system that takes the validation decision as it has been explained above.
2. A system that returns REJECTED if none token (or consecutive tokens) of the text entails some named entity in the hypothesis taking the idea of entailment described in section 2.2.

The results of the experiment are shown in table 1, showing that the second system achieves a slightly improvement in f measure, the one used for comparing AVE systems [5]. Then, the second option of validation decision was taken for English triplets.

Table 1. Comparing validation decisions.

| | **f** |
|---|---|
| **System 1** | 0.3 |
| **System 2** | 0.33 |

## 2.4 Selection decision

In AVE 2007 [4], a new measure called qa_accuracy has been proposed to compare the results of AVE participants with the results of QA participants. The objective is to measure the performance of the answer validation system selecting an answer from a set of answers to the same question. For this purpose, it is mandatory in the task that when a system returns the value VALIDATED for one or more answers to the same question, one of them has to be tagged as SELECTED.

The system we have presented does not have a way to decide what of the answers given as VALIDATED is the most probable to be correct. Then, we did not have an objective method for select answers to compare our system with the QA participants. For this reason, we decided to use a non-informative method that tagged as SELECTED the first answer of each question that is detected as correct for our system.

## 3. Results and Error Analysis

The described system has been tested in the Spanish and English test sets of AVE 2007. Tables 2 and 3 show the precision, recall and f measure over correct answers obtained respectively in English and Spanish, with a baseline system that returns VALIDATED for all the answers.

Table 2. Results in English

| | f | precision | recall |
|---|---|---|---|
| UNED system | 0.34 | 0.22 | 0.71 |
| 100% VALIDATED baseline | 0.19 | 0.11 | 1 |

Table 3. Results in Spanish

| | f | precision | recall |
|---|---|---|---|
| UNED system | 0.47 | 0.33 | 0.82 |
| 100% VALIDATED baseline | 0.37 | 0.23 | 1 |

In both languages, the results obtained have been better than the baselines, achieving a high recall.

The errors detected in triplets where the system returns VALIDATED were due to the lack of knowledge. In these pairs, all the named entities from the question and the answer are entailed for some named entity in the text snippet. However, the answer is incorrect as for example the answer in figure 3 where the expected answer is an instrument, but the given answer is a year. As the named entities of the question and the answer are entailed, our system returns VALIDATED.

**Question:** What instrument did Swann play in the duo Flanders and Swann?
**Answer:** 1964

Figure 3. Example of a false VALIDATED answer.

Regarding errors in triplets where the system returns REJECTED, in some of them a full name of a person (for example Steve Fosset) appeared in the question and the answer was judged as correct, but in the snippet appeared only the last name of this person (Fosset in the previous example). Our system cannot find a named entity in the text snippet that entails the full name and then it returns REJECTED. As it is not sure that the person in the text was the same as in the question, we think that maybe this kind of answers should be assessed as unsupported (and then in AVE as REJECTED).

Regarding the measure qa_accuracy, tables 4 and 5 show respectively the results obtained in English and Spanish, compare with the value obtained in a perfect selection and a baseline system that validates 100% of the answers and selects randomly one of them. With qa_accuracy it is also given the normalization of this measure with the perfect selection value.

Table 4. qa_accuracy results in English

| | qa_accuracy | Normalized |
|---|---|---|
| **Perfect selection** | 0.3 | 100% |
| **UNED** | 0.16 | 55% |
| **Random** | 0.1 | 35% |

Table 5. qa_accuracy results in Spanish

| | qa_accuracy | Normalized |
|---|---|---|
| **Perfect selection** | 0.59 | 100% |
| **UNED** | 0.42 | 70.3% |
| **Random** | 0.25 | 41.45% |

Even though the system uses a non-informative method for selecting answers, as it can be seen, the results are between a perfect and a random selection.

## 4. Conclusions and future work

We have presented to AVE 2007 a system based in textual entailment that does not need to build textual hypothesis. The system uses only information about named entities and obtains results very promising. These results aim us to use information about named entities in more complex answer validation systems.

We consider that the information about named entities can be used in two different ways:

1- As additional information in another answer validation system.
2- As a filter before using another answer validation system. Our system would reject answers that considers as incorrect and another system would take the decision in the rest of the answers. This idea arise from the fact that our system is focused in detecting incorrect answers achieving a precision of 95% and 90% in English and Spanish respectively over REJECTED answers.

Future work is focused in improving the named entity recognition and the decision of entailment. In this way, next step is to be able of detecting the equivalence between some named entities and their acronym (for example, UN is equivalent to United Nations).

## 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.

## References


1. Xavier Carreras, Isaac Chao, Lluís Padró and Muntsa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC04). Lisbon, Portugal, 2004.
2. V. I. Levensthein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. In Soviet Physics - Doklady, volume 10, pages 707-710, 1966.
3. 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.
4. Anselmo Peñas, Álvaro Rodrigo and Felisa Verdejo. Overview of the Answer Validation Exercise 2007. Working Notes of CLEF 2007.
5. 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.
6. Álvaro Rodrigo, Anselmo Peñas, Jesús Herrera and Felisa Verdejo, 2007. Experiments of UNED at the Third Recognising Textual Entailment Challenge Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 89-94, Prague 2007
7. Álvaro Rodrigo, Anselmo Peñas, Jesús Herrera and Felisa Verdejo, 2007. The Effect of Entity Recognition on Answer Validation. CLEF 2006, Lecture Notes in Computer Science LNCS 4730. Springer-Verlag, Berlín.