# A Supervised Learning Approach to Spanish Answer Validation

Alberto Téllez-Valero, Manuel Montes-y-Gómez, and
Luis Villaseñor-Pineda

Instituto Nacional de Astrofísica, Óptica y Electrónica
Grupo de Tecnologías del Lenguaje
Luis Enrrique Erro no. 1, Sta. María Tonantzintla, Pue.; 72840; Mexico
{albertotellezv,mmontesg,villasen}@inaoep.mx

**Abstract.** This paper describes the results of the INAOE's answer validation system evaluated at the Spanish track of the AVE 2007. The system is based on a supervised learning approach that considers two kinds of attributes. On the one hand, some attributes indicating the textual entailment between the given support text and the hypothesis constructed from the question and answer. On the other hand, some new features denoting certain answer restrictions as imposed by the question's type and format. The evaluation results were encouraging; they reached a F-measure of 53% (the best performance in the Spanish track), and outperformed the standard baseline by 15 percentage points.

## 1 Introduction

Given a question, a candidate answer and a support text, an answer validation system must decide whether accept or reject the candidate answer. In other words, it must determine if the specified answer is correct and supported.

In the previous answer validation exercise (AVE 2006), the answer validation systems were based on the idea of recognizing the textual entailment between the support text and an affirmative sentence (called hypothesis) created from the combination of the question and answer. In order to accomplish this recognition, these systems probed several approaches, ranging from simple ones taking advantage of lexical overlaps to more complexes founded on the use of a logic representation [1].

The approach based on lexical overlaps is quite simple, but surprisingly it has achieved very competitive results. Representative methods of this approach determine that H (the hypothesis) is entailed from T (the support text) only considering characteristics such as named entity overlaps [2], n-gram overlaps and the size of the longest common subsequence (LCS) [3].

The simplicity is the strength of this approach but at the same time is its weakness. All overlap-based methods have problems to deal with situations where the answer should be satisfy simple type restrictions imposed by the question. For instance, the candidate answer *"Javier Sotomayor"* is clearly incorrect for the question *"What is the world record in the high jump?"*, but it will be

validated as accepted because the high lexical similarity between the formed hypothesis *"The world record in the high jump is Javier Sotomayor"* and the corresponding support text *"The world record in the high jump, obtained by Javier Sotomayor, is 2.45 meters."*.

The proposed system adopts several ideas from recent systems (in particular from [2, 3]): it is based on a supervised learning approach that considers a combination of some previously-used features. However, in addition, it also includes some new characteristics that allow tackling the discussed problem.

## 2 The Answer Validation System

In resume, the main characteristics of our system are the following:

1. It only considers content words for computing word overlaps and LCS.
2. It uses POS tags for the calculus of the LCS.
3. It makes a syntactic transformation of the generated hypothesis in order to simulate the active and passive voices.
4. It applies some manually-constructed lexical patterns to help treating support texts containing an apposition and adjectival phrases.
5. It includes some new features denoting certain answer restrictions as imposed by the question's class.

For a complete description of the system refer to [4].

## 3 Experimental Evaluation

### 3.1 Training and Test Sets

In order to avoid the low recall in the validated answers we assembled a more balanced training set. Basically, we joined some answers from the training sets of the AVE 2006 and 2007. This new training set contains 2022 answers, where 44% are validated and 56% rejected. On the other hand, the evaluation set for the Spanish AVE 2007 contains 564 answers (22.5% validated and 77.5% rejected) corresponding to 170 different questions.

### 3.2 Results

This year we submitted two different runs considering two different classification algorithms. The first run (RUN 1) used a single support vector machine classifier, whereas the second run (RUN 2) employed an ensemble of this classifier based on the AdaBoostM1 algorithm.

Table 1 shows the evaluation results corresponding to our two submitted runs. It also shows (in the last row) the results for a 100% VALIDATED baseline (i.e., an answer validation system that accepted all given answers). The results indicate that our methods achieved a high recall and a middle level precision,

which means that they correctly accepts must of the right answers (there are a few false negatives), but also incorrectly accepts some wrong ones (there are several false positives).

An analysis of false positives shows us that the main problem of our approach is still the high overlap that exists between the T and H although the evaluated answer is wrong. For instance, in the question *"Who made Windows 95?"*, the wrong candidate answer *"business"* is validated as accepted. This error occurs because the content terms in the formed hypothesis *"business made Windows 95"* can be totally overlap by the support text *"Windows 95 is the new version of the operating system made for the business Microsoft, ..."*. These cases evidenced the necessity of including more information into the overlap checking process, such as term dependencies and more restrictive data about the kind of expected answer.

**Table 1.** General evaluation of the INAOE's system (here TP, FP, TN, and FN refers to true positives, false positives, true negatives, and false negatives, respectively)

| | TP | FP | TN | FN | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|---|
| RUN 1 | 109 | 176 | 248 | 18 | 0.38 | 0.86 | 0.53 |
| RUN 2 | 91 | 131 | 293 | 36 | 0.41 | 0.72 | 0.52 |
| 100% VALIDATED | 127 | 424 | – | – | 0.23 | 1 | 0.37 |

This year the AVE organizers decide to include a new evaluation measure, called qa-accuracy. This measure allows evaluating the influence of the answer validation systems into the question answering task. In order to compute this measure the answer validation systems must select only one validated answer for each question. This way, the qa-accuracy expresses the rate of correct selected answers. Table 2 presents the qa-accuracy results of our two runs. It also shows (in the last row) the best results obtained at QA@CLEF 2007 for the same set of questions.

**Table 2.** Evaluation results obtained by the qa-accuracy measure

| | Selected answers | | | | QA-accuracy |
|---|---|---|---|---|---|
| | Total | Right | Wrong | Inexact | |
| RUN 1 | 129 | 76 | 47 | 6 | 0.45 |
| RUN 2 | 107 | 62 | 40 | 5 | 0.36 |
| BEST QA SYSTEM | – | 84 | – | – | 0.49 |

In order to do a detail evaluation of our system we also measured its precision over the subset of 101 questions that have at least one correct candidate answer. In this case, RUN 1 selected the right candidate answer for 75% of the questions,

and RUN 2 for 61%. For the rest of the questions (69 questions), for which no correct candidate answer exists, RUN 1 correctly answered NIL in 49% of the cases, whereas the RUN 2 correctly responded NIL in 61% of the questions.

It is important to mention that current qa-accuracy measure does not take into account the correctly selected NIL answers. That is, it does not consider NIL answers as correct answers for any question (even for those cases that do not have the answer in the test document collection). Considering NIL answers into the evaluation, our answer validation system – in the RUN 1 – could reach an accuracy equal to the best QA system (i.e., 49%).

## 4 Conclusions

This paper presents the evaluation results of the INAOE's answer validation system at the Spanish track of the AVE 2007. Our system adopts several ideas from recent overlap-based methods; basically, it is based on a supervised learning approach that uses a combination of some previous used features, in particular, word overlaps and longest common subsequences. However, it also includes some new notions that extend and improve these previous methods.

The evaluation results are encouraging; they show that the proposed system achieved a 53% of F-measure, obtaining the best result in the Spanish track. As future work we plan to enhance the question-answer compatibility analysis as well as to apply other attributes in the supervised learning process.

### Acknowledgments

This work was done under partial support of CONACYT (project grant 43990 and scholarship 171610). We also thank the CLEF organizers.

## References

1. Peñas, A., Rodrigo, Á., Sama, V., Verdejo, F.: Overview of the answer validation exercise 2006. [5] 257–264
2. Rodrigo, Á., Peñas, A., Herrera, J., Verdejo, F.: The effect of entity recognition on answer validation. [5] 483–489
3. Kozareva, Z., Vázquez, S., Montoyo, A.: University of Alicante at QA@CLEF2006: Answer validation exercise. [5] 522–525
4. Téllez-Valero, A., Montes-y-Gómez, M., Villaseñor-Pineda, L.: INAOE at AVE 2007: Experiments in Spanish answer validation. In: Working notes for the 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary (September 19-21 2007)
5. Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M., eds.: Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers. In Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M., eds.: CLEF. Volume 4730 of Lecture Notes in Computer Science., Springer (2007)