INAOE at AVE 2007: Experiments in Spanish Answer Validation

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

Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico.
{albertotellezv, mmontesg, villasen}@inaoep.mx

Abstract

This paper describes the INAOE's answer validation system evaluated at the Spanish track of the AVE 2007. This 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. In order to extract all these attributes the system uses different tools such as a lemmatizer, a POS tagger, a NER procedure and a superficial syntactic parser. Experimental results are encouraging; they show that the proposed system achieved a 52.91% of F-measure and that it outperformed the standard baseline by 15 percentage points.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [Database Management]: Languages—Query Languages

General Terms

Measurement, Performance, Experimentation

Keywords

Answer Validation, Question Answering, Textual Entailment Recognition, and Supervised Learning.

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.

Answer validation systems have been traditionally 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 the answer. In order to accomplish this recognition they have probed several approaches, ranging from simple ones taking advantage of lexical overlaps to more complexes founded on the use of a logic representation [6].

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 [8], n-gram overlaps [4], as well as the size of the longest common subsequence (LCS) [1,5].

The simplicity is the strength of this approach, but also it is its weakness. For instance, [8] can easily recognize the textual entailment between "*Lucy visits some friends*" (H) and "*Lucy goes with some friends*" (T). However, it fails when there is a high word overlap between H and T, but there does not exist an entailment relation; for example between H and the text "*Lucy has some wonderful friends*". The method presented in [4] has the same problem. Nevertheless it is much restrictive on evaluating the overlap between H and T, and therefore it tends to produce better results.

The methods based on the use of LCSs [1,5] are less sensible to the word overlap rate, but they are still very sensible to changes in the word order (like in the use of active and passive voices). For instance they will be unsuccessful in recognizing the entailment between "*Lucy visits some friends*" (H) and "*Some friends are visited by Lucy*" (T).

Finally, all overlap-based methods have problems to deal with situations where the answer does not satisfied simple type restrictions imposed by the question. For instance, in the example of Table 1, the candidate answer is clearly incorrect, but it will be validated because the high lexical similarity between H and T.

Table 1. Incorrect answer validation using overlap-based methods

Question:	What is the world record in the high jump?
Answer:	Javier Sotomayor
Support text (T):	The world record in the high jump, obtained by Javier Sotomayor, is 2.45 meters.
Hypothesis (H):	Javier Sotomayor is the world record in the high jump

The system described in this paper adopts several ideas from recent systems (in particular from [1,4,5]). It is based on a supervised learning approach that uses a combination of all previous used features (word overlaps and LCSs). In addition, it also includes some new characteristics that allow reducing previously discussed problems. In particular:

- It considers only content words for the calculus of word overlaps and the LCS. This represents a middle point between the usage of all words [4] and just named entities [8].
- It computes the LCS taking into consideration POS tags. This characteristic makes possible obtaining larger subsequences and also dealing with the synonymy phenomena.
- It makes a simple syntactic transformation over the generated hypothesis in order to simulate the usage of active and passive voices. In other words, it generates two hypotheses (H and H´) that combine in a different way the given question and answer. The inclusion of this additional hypothesis improves in many cases the calculus of the LCS.
- It uses some manually constructed lexical patterns to help treating support texts containing an apposition phrase. This idea was taken from the COGEX logic-based system [9]. Its goal is to make explicit the relation between the two elements of the apposition phase.
- Finally, it includes some new features denoting certain answer restrictions as imposed by the question's class. This new features avoid validating answers that does not correspond to the expected semantic type of answer.

The following sections give some details on the proposed system. In particular, section 2 describes the main characteristics of our system, whereas section 3 presents our evaluation results on the AVE 2007 Spanish track. Finally, section 4 discusses some general conclusions about the performance of the proposed approach.

2 Our Answer Validation System

Figure 1 shows the general architecture of our system. It consists of three main phases: preprocessing, feature extraction and classification. The following subsections describe these processes.

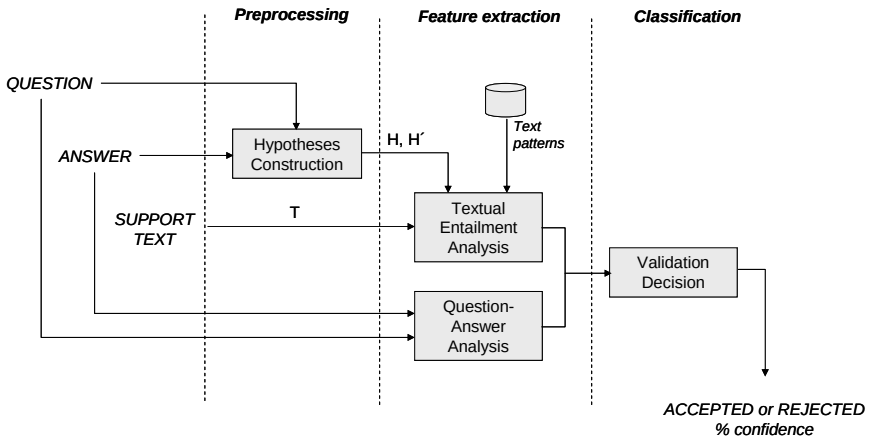

Figure 1. INAOE's system for Spanish answer validation

2.1 Preprocessing

The main task of this initial phase is to construct *two distinct hypotheses* combining the given question and answer. In order to do that it firstly applies a superficial syntactic analysis over the question[1]. Then, using the obtained syntactic tree, it generates both hypotheses.

The first hypothesis (H) is constructed by replacing the nominal phrase that contains the interrogative particle by the given answer. For instance, given the question "*How many inhabitants are there in Longyearbyen?*" and the answer "*180 millions of inhabitants*", this approach allows generating the hypothesis H = "*180 millions of inhabitants are there in Longyearbyen*".

The second hypothesis (H´) is obtained doing a simple transformation on H. The idea is to detect the main verb phrase of the H (that is the main verb phrase of the question) and then interchange its surrounding nominal phrases. This way the second hypothesis for our example is H´ = "*in Longyearbyen are there 180 millions of inhabitants*".

2.2 Feature Extraction

As we previously mentioned, our system considers two kinds of attributes. On the one hand, some attributes indicating the textual entailment between the support text and the constructed hypotheses. On the other hand, some new features that denote certain answer restrictions imposed by the question's type. These attributes are extracted by two different modules of our system, the textual entailment analysis and the question-answer analysis.

Textual Entailment Analysis

The textual entailment analysis of the pairs (T, H) and (T, H´) consists of two stages: (i) compute the *term overlap*, and (ii) calculate the *term sequence overlap*. In order to avoid a high matching caused by functional terms (such as prepositions and determiners), in both cases we only consider the occurrence of content terms (nouns, verbs, adjectives and adverbs). Besides, we use the word lemmas, which allow getting a better term overlap.

The term overlap between the pair (T, H) is computed by a simple counting of the common content words in the support text (T) and the hypothesis (H)[2]. The following features are generated from this analysis:

(1) The rate of noun overlap between (T, H)
(2) The rate of verb overlap between (T, H)
(3) The rate of adjective overlap between (T, H)
(4) The rate of adverb overlap between (T, H)
(5) The rate of date overlap between (T, H)
(6) The rate of number overlap between (T, H)

Similar to [1,5] we compute the term sequence overlap by extracting the longest common subsequence (LCS). However, different from these previous approaches, our method only considers the occurrence of content words and allows the inclusion of POS tags inside the sequence.

In this case, it is necessary to compute the LCS from (T, H) as well as from (T, H´). Nevertheless, only the longest subsequence is used. This way we generate the following feature from this analysis:

(7) The size of the LCS between (T, H) or (T, H´) divided by the size of H

It is important to remember that the presence of apposition phrases in the support text causes detriment on the LCS. In order to solve this problem we propose using some manually-constructed transformation patterns such as "*The <AGENT>, <AGENT>, → The <AGENT> <V> <AGENT>*".

The example of Table 2 illustrates the application of these patterns as well as the inclusion of POS tags inside the LCSs.

Question-Answer Analysis

There are two common situations related with the presence of an incorrect answer. The first one is that the semantic class of the extracted answer does not correspond to the expected class of answer (in accordance to the given question). For instance, having the answer "*yesterday*" for the question "*How many inhabitants are there in Longyearbyen?*".

The second situation occurs when the question asks about a specific fact and the answer makes reference to another different one. For instance, answering "*eight*" to the example question, using as support text "…*when eight animals parade by the principal street in Longyearbyen, a town of a thousand of inhabitants*".

[1] This analysis was done by Freeling [2], an open source suite of language analyzers.
[2] It is not necessary to compute the term overlap between (T) and (H´) since it will be exactly the same.

Table 2. Application of transformation patterns and POS tags in the LCS calculus

Question:	What is the quinua?
Answer:	Cereal
Support text (T):	The quinua, an American cereal of great nutritional value, …
Original analysis	
Hypotheses:	Cereal is the quinua (H), The quinua is cereal (H´)
LCS (of size = 2):	Quinua cereal
Using the transformation patterns	
Support Text (T´):	The quinua (V) an American cereal of great nutritional value …
Hypotheses:	Cereal is the quinua (H), The quinua is cereal (H´)
LCS (of size = 3):	Quinua (V) cereal

Our system includes two new features that attempt to capture these situations:

(8) A Boolean value indicating if a general-class restriction is satisfied, and
(9) A Boolean value indicating if a specific-type restriction is satisfied

The *general-class answer restriction* is TRUE if the semantic class of the extracted answer and the expected class of the answer are equal; other case it is set to FALSE. We consider three general classes: quantity, date, and proper noun. The question classification (i.e., the definition of the expected class of the answer) is done using the KNN supervised algorithm[3] with K = 1.

In order to determine the *specific target fact* concerning the question it is necessary to perform the following procedure: (i) construct the syntactic tree of the question, and (ii) extract the principal noun from the noun phrase that contains the interrogative particle. Applying this procedure over the example question, the word "*inhabitants*" was selected as the specific target fact.

Once extracted the specific target fact from the question, it is possible to evaluate the *specific-type answer restriction*. Its value is set to TRUE if the specific target fact happens in the support text, in the immediate answer context (one content word to the right or left). In any other case its value is set to FALSE. Therefore, the candidate answer "*eight*" has its value set to FALSE since its immediate context ("eight *animals*") does not contains the noun "*inhabitants*". On the contrary, the candidate answer "*thousand*" will be have its value set to TRUE, since the noun "*inhabitants*" occurs in its immediate context ("*town* thousand *inhabitants*").

It is important to notice that not for all questions it is possible to establish a specific target fact (e.g., consider the question "*When was Amintore Fanfani born?*"). In these cases we considered –by default– that all candidate answers satisfied the specific-type restriction.

2.3 Classification

This final module generates the answer validation decision by means of a supervised learning approach, in particular, by a support vector machine classifier. This classifier decides if the answer is *validated* or *rejected* on the basis of the nine previously described features along with the following two additional ones:

(10) The question category (i.e., factoid, definition, or list)
(11) The question interrogative particle (i.e., who, where, when, etc.)

An evaluation of the proposed features during the development phase, using the information gain algorithm, shows us that the nouns overlap and the LCS size are the most discriminative features. The general ranking of the eleven features in decreasing order is as follows: 1, 7, 6, 11, 10, 8, 2, 5, 9, 4, and 3.

3 Experimental Evaluation

3.1 Training and Test Sets

The training set available for the AVE 2007 Spanish task consists of 1817 answers, where 15% are validated answers and the rest 85% are rejected. 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.

Details on these sets are described in [7].

[3] For the training process we considered all questions from the previous question answering CLEF campaigns.

3.2 Results

This section describes the experimental results of our participation at the AVE 2007 Spanish task. This year we submitted two different runs. The first run (RUN 1) considered the system just as it was described in the previous section. On the other hand, the second run (RUN 2) used a different learning method; instead of using a single support vector machine classifier, it employed an ensemble of this classifier. This ensemble was implemented using the AdaBoostM1 algorithm in Weka [3].

Table 3 shows the evaluation results corresponding to our two submitted runs. It also shows (in the last row) the results for a 100% YES baseline (i.e., an answer validation system that validated all given answers). The results indicate that our methods achieved a very high recall and a middle level precision, which means that it validates must of the correct answers, but also some incorrect ones. It is important to point out that our best result (RUN 1) outperformed the baseline by 15 percentage points; the same proportion than the best evaluated system at the AVE 2006 [6].

Table 3. General evaluation of the INAOE's system

	TP	FP	TN	FN	Precision	Recall	F-measure
RUN 1	109	176	248	18	38.25%	**85.83%**	**52.91%**
RUN 2	91	131	293	36	**40.99%**	71.65%	52.15%
100% YES	127	424	-	-	23.05%	100%	37.46%

This year the AVE organizers decide to include a new evaluation measure. This new measure, called qa-accuracy, aims to evaluate the influence of the answer validation systems to the question answering task. In order to compute this measure the answer validation systems must to select only one validated answer for each question. This way, the qa-accuracy expresses the rate of correct selected answers.

Table 4 presents the qa-accuracy results of our two runs. It also shows the results obtained by an "ideal" answer validation system (i.e., a system that, when possible, always selects a correct answer). Here, it is necessary to clarify that because only 101 questions (from the whole set of 170) has a correct candidate answer, it is impossible to obtain a 100% qa-accuracy.

The results of Table 4 are not conclusive. However, it is interesting to comment that our QA system (that was the best one in 2005 and the second best one in 2006 in the Spanish QA task) [10] obtains a 35.88% of accuracy on the same questions set and ignoring evaluate the correct NIL questions. This fact indicates –in some way– that answer validation is useful, and that it could produce interesting improvements over current QA systems.

Table 4. Evaluation results obtained by the qa-accuracy measure

	Selected Answers				
	Total	Right	Wrong	Inexact	QA-accuracy
RUN 1	129	76	47	6	44.71%
RUN 2	107	62	40	5	36.47%
IDEAL	101	101	-	-	59.41%

In order to do a detail evaluation of our system we also measured its precision over the subset of 101 questions that have a candidate corrected answer. In this case, RUN 1 validated the correct candidate answer for 75% of the questions, and RUN 2 for 61%. For the rest of the questions (a subset of 69 questions), where does not exist any correct candidate answer, the RUN 1 correctly answered NIL in 49% of the cases, whereas the RUN 2 correctly responded NIL in 61% of the questions.

4 Conclusions

This paper described the INAOE's answer validation system that was evaluated at the Spanish track of the AVE 2007. This system adopts several ideas from recent *overlap-based* methods; basically, it is based on a supervised learning approach that uses a combination of all previous used features, in particular, the word overlaps and the longest common subsequences. However it includes some new notions that extend and improve these previous methods. For instance: (i) it considers only content words for the calculus of word overlaps and the LCS; (ii) it computes the LCS taking into consideration POS tags; (iii) it makes a syntactic transformation over the generated hypothesis in order to simulate the active and passive voices; (iv) it uses some manually constructed lexical

patterns to help treating support texts containing an apposition phrase; (v) it includes some new features denoting certain answer restrictions as imposed by the question's class.

The evaluation results are encouraging; they show that the proposed system achieved a 52.91% of F-measure and that it outperformed the standard baseline by 15 percentage points. Moreover, they indicate that our system is especially precise (75% of accuracy) in selecting the correct answer for a question when such answer exists inside the set of candidate answers.

Acknowledgements. This work was done under partial support of CONACYT (project grant 43990 and scholarship 171610). We also like to thanks to the CLEF organizing committee as well as to the EFE agency for the resources provided.

References

1. Bosma W., and Callison-Burch C. *Paraphrase Substitution for Recognizing Textual Entailment*, In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006), Alicante, España, September 2006.
2. Carreras, X., I. Chao, L. Padró and M. Padró. *FreeLing: An Open-Source Suite of Language Analyzers*, Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal. 2004.
3. Freund Y., and Schapire R. *Experiments with a new boosting algorithm*, Proc International Conference on Machine Learning, pages 148-156, Morgan Kaufmann, San Francisco, 1996.
4. Herrera J., Rodrigo A., Peñas A., and Verdejo F. *UNED Submission to AVE 2006*, In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006), Alicante, España, September 2006.
5. Kozareva Z., Vázquez S., and Montoyo A. *Adaptation of a Machine-learning Textual Entailment System to a Multilingual Answer Validation Exercise,* In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006), Alicante, España, September 2006.
6. Peñas A., Rodrigo A., Sama V., and Verdejo F. *Overview of the Answer Validation Exercise 2006,* In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006), Alicante, España, September 2006.
7. Peñas A., Rodrigo A., Sama V., and Verdejo F. *Overview of the Answer Validation Exercise 2007,* In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2007), Budapest, Hungary, September 2007. *In this volume.*
8. Rodrigo A., Peñas A., and Verdejo F. *The Effect of Entity Recognition in Answer Validation,* In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006), Alicante, España, September 2006.
9. Tatu M., Iles B., and Moldovan D. *Automatic Answer Validation using COGEX,* In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006), Alicante, España, September 2006.
10. Téllez A., Juárez A., Hernández G., Delicia C., Villatoro E., Montes M., and Villaseñor L. *INAOE's Participation at QA@CLEF 2007*, In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2007), Budapest, Hungary, September 2007. *In this volume.*