Analyzing the Use of Non–Overlap Features for Supervised Answer Validation

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

Laboratory of Language Technologies, INAOE, Mexico
{albertotellezv,antjug,mmontesg,villasen}@inaoep.mx

Abstract. This year we evaluated our supervised answer validation method at both, the Spanish Answer Validation Exercise (AVE) and the Spanish Question Answering Main Task. This paper describes and analyzes our evaluation results from both tracks. In resume, the F-measure of the proposed method outperformed the baseline result of the AVE 2008 task by more than 100%, and enhanced the performance of our question answering system, showing a gain in accuracy of 22% for answering factoid questions. A detailed analysis of the results shows that the proposed non–overlap features are most discriminative than the traditional overlap ones. In particular, these novel features allowed increasing the F-measure result of our method by 26%.

1 Introduction

An *answer validation* (AV) method try to determine if a specified answer is correct and supported. These methods are especially useful for filtering the best responses from question answering (QA) systems and for superficially combining them. In line with these efforts, we implemented a new AV method based on a supervised learning approach. In particular, our method implements a boosting ensemble —formed by ten decision tree classifiers— that decides whether to accept or reject each candidate answer based on the use of ninety-six attributes that characterize: (*i*) the compatibility between question and answer types; (*ii*) the redundancy of answers across streams; and (*iii*) the overlap and the non-overlap between the question-answer pair and the core fragment of the support text.

In order to evaluate the proposed method we considered two different scenarios: the Answer Validation Exercise (AVE 2008) and the Question Answering Main Task (QA@CLEF 2008). The objective of the first scenario was to evaluate the ability of our AV method to discriminate correct from incorrect answers as well as its capacity to combine the answers from several QA systems. In contrast, the goal of the second evaluation scenario was to measure the impact of including an answer validation module in our QA system [1].

The evaluation results were encouraging; the proposed method outperformed the F-measure result of the baseline at AVE 2008 task by more than 100%, and also enhanced the performance of our QA system producing a gain in accuracy of

22% for factoid questions. The analysis of the results showed the importance of the proposed non-overlap features, which increased our F-measure in 26%. This analysis also indicated that the redundancy features were not useful for the AVE 2008 test set; their elimination allowed achieving a gain in F-measure of 5%.

Due to space limitations, we decide to omit the description of our AV method (which can be found in [2]), and exclusively consider the analysis of the evaluation results. In particular, this paper is organized as follows. Section 2 resumes the results of the proposed method at CLEF. Section 3 presents the results analysis, and Section 4 exposes our conclusions and outlines some future work directions.

2 Evaluation Results

2.1 Spanish Answer Validation Exercise

Table 1 shows the answer validation results corresponding to our two submitted runs to Spanish AVE 2008 [3]. It also shows the results for the baseline (a 100% VALIDATED). The results indicate that relaxing the acceptance threshold over the answers's confidence value (RUN 1) our method achieved a high recall but a low precision. In contrast, the second run that maintained the default threshold (RUN 2) got a worst recall, but achieved a major precision overcoming the baseline F-measure result in more than 100%.

Table 1. Results for the answer validation evaluation

	Precision	Recall	F-measure
RUN 1	0.13	0.86	0.23
RUN 2	0.30	0.59	0.39
100% VALIDATED	0.10	1.0	0.18

Complementary to the previous data, Table 2 shows the evaluation results for combining the answers from several QA systems. These results indicate that the QA-accuracy of RUN 1 is 19% better than the result of RUN 2. Given that RUN 2 outperformed the answer validation result of RUN 1 (see Table 1), these results confirm our observation in [2] that the best answer validation method (but not perfect) not necessary produces the best QA stream fusion performance. The results in Table 2 also indicate that, because of the better capacity of RUN 2 to rejected wrong answers, the estimated-QA-performance of both runs were very similar. Refer to [3] for a description of the evaluation measures at AVE.

2.2 Spanish Question Answering Main Task

This year we submitted two different runs at the Main QA task [4]. The first run (inao081eses) was the original output of our QA system (refer to [1] for details), whereas the second run (inao082eses) was the result of applied the AV method

Table 2. Results for the QA stream fusion evaluation

	QA-accuracy	QA-reject-accuracy	estimated-QA-performance
RUN 1	0.32	0.06	0.34
RUN 2	0.27	0.22	0.33
PERFECT FUSION	0.62	0.38	0.85

over the set of candidate answers generated by the first run. Table 3 shows the evaluation results of both runs as well as a baseline result corresponding to a perfect validation of the output of our QA system.

Table 3. Results of the QA main task (It shows the accuracy, as well as the number of questions answered right (R), wrong (W), inexact (X), and unsupported (U))

	Factual				Definition				
	R	W	X	U	R	W	X	U	Accuracy
inaoe081eses (original QA system)	23	156	1	1	19	0	0	0	0.21
inaoe082eses (QA system with an AV module)	28	149	3	1	16	3	0	0	0.22
PERFECT VALIDATION	30	147	3	1	19	0	0	0	0.25

Results from Table 3 indicate that the AV module helped increasing the number of right answers for factoid questions, improving the accuracy of our QA system by 22%. In contrast, the AV module damaged the treatment of definition question since it incorrectly rejected three right answers.

3 Results Analysis

In order to understand the behavior of the proposed AV method, we carried out a deep analysis of the usefulness of each one of the used characteristics over the AVE 2008 test set. The information gain (IG) values for the features used by our supervised AV method show two interesting facts. First, the proposed non–overlap features —with an average IG of 0.024— were more discriminative than the traditional overlap features (which got an average IG of 0.004). And second, in contrast to the high IG of the answer redundancy feature in the train set (a 0.184), in the test set this feature only reached a 0.015 of IG.

To evaluate the effects of these two facts over the performance of our AV method, we run two extra experiments: *i*) eliminating the attributes related to the non–overlap features; and *ii*) eliminating the attribute related to the answer redundancy. Taking as baseline the F-measure of our RUN 2 (see Table 2), the results of these extra experiments showed the following. First, the elimination of the non–overlap features decreased our baseline result in a 26% (getting a F-measure of 0.29), this result indicated the importance of these characteristics

for answer validation. Second, the elimination of the answer redundancy feature allowed improving the baseline result by 5% (reaching a F-measure of 0.41), indicating that it is not relevant for this particular data set.

In relation to include the AV module into our QA system, the results in Table 3 indicate that it was only useful for factoid questions. However, due to the small number of extracted right answers for this kind of questions, it was impossible to obtain a better improvement. This fact was particularly evident for the questions answered from Wikipedia as described in [1]. Finally, taking into account that our QA system is very accurate for answering definition questions, we conclude that the use of an AV module is not convenient for this kind of questions.

4 Conclusions

This paper showed and discussed the evaluation results of our supervised AV system. The obtained results at two different scenarios (the Spanish AVE 2008 and the Spanish QA@CLEF 2008) were encouraging; the proposed method achieved a F-measure of 0.39 in the detection of correct answers, outperforming the baseline result by more than 100%. It also enhanced the performance of our Spanish QA system, producing a gain in accuracy of 22% for factoid questions. An analysis about of the results showed that the proposed non-overlap characteristics are more discriminative than the traditional overlap features, contributing in a 26% of the F-measure result.

Finally, it is important to comment that this year our best results in the AVE (a F-measure of 0.39 and a QA-accuracy of 0.32) were very distant from those corresponding to a perfect validation. We presume that this situation was caused by the decreasing number of right answers together with the increasing number of relevant support passages related to the wrong answers. In order to tackle these problems, and based on the fact that non-overlap attributes were the most discriminative, we plan to include more elements (such as prepositions, conjunctions, and some punctuation marks) for their computation.

Acknowledgments

This work was done under partial support of CONACYT (project grant 61335, and scholarships 171610 and 165499).

References

1. Téllez-Valero, A., et al.: INAOE's participation at QA@CLEF 2007. In: CLEF 2007 Working Notes, Budapest, Hungary (2007)
2. Téllez-Valero, A., et al.: INAOE at QA@CLEF 2008: Evaluating answer validation in spanish question answering. In: CLEF 2008 Working Notes, Denmark (2008)
3. Rodrigo, Á., Peñas, A., Verdejo, F.: Overview of the answer validation exercise 2008. In: CLEF 2008 Working Notes, Aarhus, Denmark (2008)
4. Forner, P., et al.: Overview of the CLEF 2008 multilingual question answering track. In: CLEF 2008 Working Notes, Aarhus, Denmark (2008)