UNED Submission to AVE 2006

Jesús Herrera, Álvaro Rodrigo, Anselmo Peñas, Felisa Verdejo
Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
Madrid, Spain

{jesus.herrera, alvarory, anselmo, felisa}@lsi.uned.es

Abstract

This paper reports the participation of the Spanish Distance Learning University (UNED) in the First Answer Validation Exercise (AVE) celebrated within the Cross Language Evaluation Forum (CLEF) 2006 edition. The system works for the Spanish language. It is based on a Support Vector Machine (SVM) classification of the pairs <text, hypothesis> given by the organization. This classification is accomplished by means of a set of features obtained from lexical analysis. Yet Another Learning Environment (Yale 3.0) was used for the SVM classification. Freeling was the toolkit elected for lemmatization and named entities recognition. Two runs were submitted and the results obtained, as defined by the organizers, were the following: $precision_{run1} = 0.467$, $recall_{run1} = 0.7168$, $F_{run1} = 0.5655$, $precision_{run2} = 0.4652$, $recall_{run2} = 0.7079$, $F_{run2} = 0.5615$.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: I.2.7 Natural Language Processing: Text analysis – Language parsing and understanding

Keywords

Question Answering, Answer Validation, Textual Entailment, Entity Recognition

1. Introduction

The system presented to the First AVE is based on the ones developed for the First [4] and the Second [5] Recognizing Textual Entailment (RTE) Challenges. This is because the parallelism between both exercises. The AVE was defined in the way that a given answer to a question must be validated by means of the information contained in the question, the answer and the text supporting the answer. This information was elaborated by the organization in order to present it in the form of a pair of texts, namely text and hypothesis. The objective of the exercise is to automatically determine if one of the snippets – the text – entails the other one – the hypothesis –; if so, it is said that the answer given to the question is validated considering the information given by the supporting text. For the RTE Challenge, the participant systems must determine the existence of entailment between pairs of texts [1][2]. Thus, the systems participating in the RTE Challenge should be able to participate in the AVE. In the system here described, the basic ideas from the ones presented to the RTE Challenges were kept, but the new system was designed and developed according to the available resources for the Spanish language, lacking some subsystems implemented in the ones cited above such, for example, dependency analysis. In short, the techniques involved in this new system are the following:

- Ratio of coincidence between words, unigrams, bigrams and trigrams, respectively, from the texts and their correspondant hypothesises.
- Detection of entailment between numeric expressions of the texts and the hypothesises.
- Detection of entailment between named entities of the texts and the hypothesises.
- Support Vector Machine classification in order to determine the final decision about textual entailment between pairs of text and hypothesis.

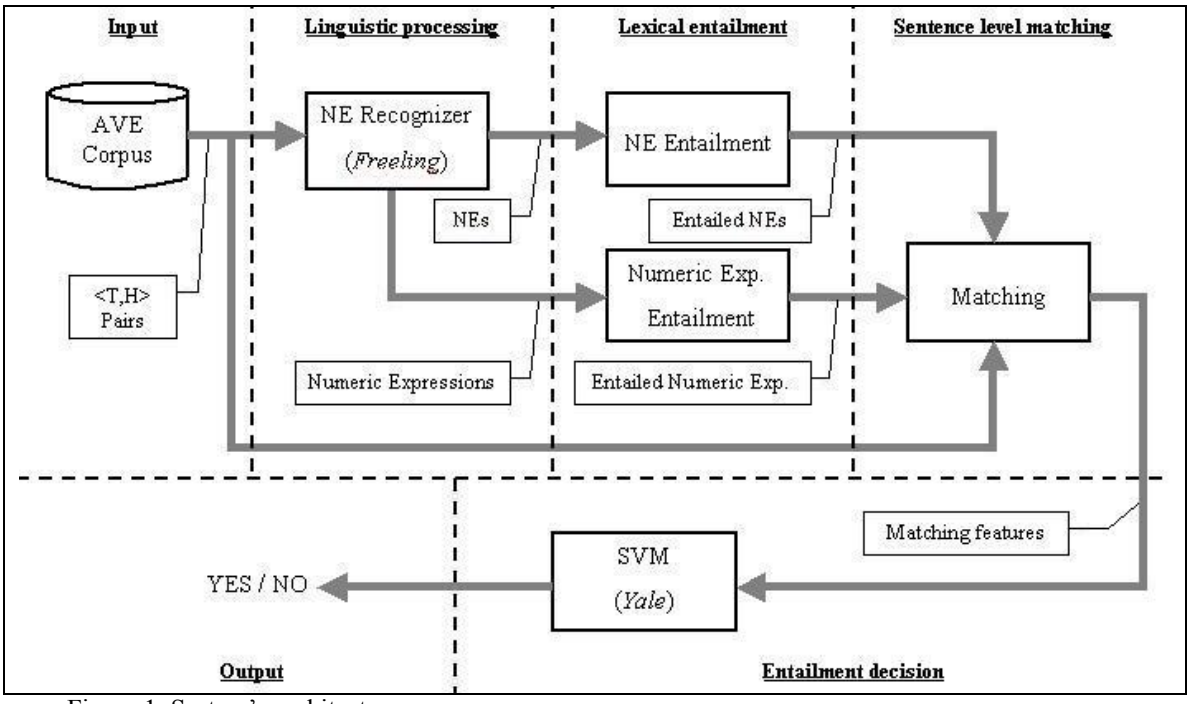

Figure 1. System's architecture

2. System description

The proposed system is based on surface techniques of lexical analysis, i.e., lemmatization and recognition of named entities. The system accepts pairs of text snippets (text and hypothesis) at the input and gives a boolean value at the output: YES if the text entails the hypothesis and NO otherwise. This value is obtained by the application of a learned model by a SVM classifier. System's components, whose graphic representation is shown in figure 1, are the following:

2.1. Linguistic processing

A named entities recognition was accomplished, using Freeling[1], in order to detect numeric expressions and named entities of the texts and hypothesises. In addition, the lemmas of every text and hypothesis were obtained using Freeling, too.

2.2. Lexical entailment

The entailment module takes the information given by the entities recognizer and computes the following features:

- Entailment between numeric expressions. Numeric expressions from the corpus are detected by means of an entities recognizer. Thus, a numeric expression N1 entails a numeric expression N2 if the range associated to N2 encloses the range of N1. When a numeric expression in the hypothesis is not entailed by one or more numeric expressions in the text, then the system responses that there is not entailment between numeric expressions in the pair. Also, the module computes the ratio of numeric expressions from every hypothesis entailed by any numeric expression from its text.

- Entailment between named entities. Named entities from the corpus are detected by means of an entities recognizer. Thus, a named entity NE1 entails a named entity NE2 if NE2 is contained in N1. When a named entity in the hypothesis is not entailed by one or more named entities in the text, then the system responses that there is not entailment between named entities in the pair. Also, the module computes the ratio of named entities from every hypothesis entailed by any named entity from its text.

2.3. Sentence level matching

[1] http://www.lsi.upc.edu/~nlp/freeling/

A plain text matching module that calculates the percentage of words, unigrams (lemmas), bigrams (lemmas) and trigrams (lemmas), respectively, from the hypothesis entailed by lexical units (words or n-grams) from the text, considering them as bags of lexical units.

2.4. Entailment decision

A SVM classifier, from Yet Another Learning Environment (Yale 3.0) [3], was applied in order to train a model from the development corpus given by the organization and to apply it to the test corpus. The model was trained by means of a set of features obtained from the other modules of the system; these ones, for every pair <text, hypothesis>, are the following:

1. Percentage of words of the hypothesis in the text (treated as bags of words).
2. Percentage of unigrams (lemmas) of the hypothesis in the text (treated as bags of unigrams).
3. Percentage of bigrams (lemmas) of the hypothesis in the text (treated as bags of bigrams).
4. Percentage of trigrams (lemmas) of the hypothesis in the text (treated as bags of trigrams).
5. Existence or absence of any numeric expression within the hypothesis entailed by any numeric expression within the text.
6. Existence or absence of any named entity within the hypothesis entailed by any named entity within the text.
7. Ratio of numeric expressions within the hypothesis entailed by any numeric expression within the text.
8. Ratio of named entities within the hypothesis entailed by any named entity within the text.

3. Description of the runs submitted

Two runs were submitted to the First AVE.

Run 1 was obtained using all the features described in section 2.4, computed from the development corpus, to train a model with the SVM.

Run 2 was obtained using only the features 1, 2, 3 and 4 described in section 2.4, computed from the development corpus, to train a model with the SVM.

The reason that motivated the election of features for run 1 and run 2 was to show the importance of considering named entities and numeric expressions for this kind of exercises, as described in [5].

Thus, both models were applied to the appropriate features calculated from the test corpus in order to obtain a pair of predictions (run 1 and run 2) for the existence or absence of textual entailment for every pair <text, hypothesis>.

4. Results plus analysis of the results

Three evaluation measures were applied to the participating systems: precision (the ratio between the number of pairs correctly labeled with YES as the predicted value for the entailment and the total number of pairs labeled with YES as the predicted value), recall (the ratio between the number of pairs correctly labeled with YES as the predicted value for the entailment and the total number of pairs having YES as their real value of entailment) and F-measure (giving the same weight to precision and recall).

While the development time, two models were obtained by training the SVM: one of them using all the features described in section 2.4 (such as run 1) and the other one using only the features 1, 2, 3 and 4 described in section 2.4 (such as run 2). The model was trained by means of a part of the development corpus and tested by means of the other part or that corpus. For this experiment, the evaluation measures above mentioned show the values below.

- Run 1 (development):
    - precision = 0.6431

- recall = 0.8967
- F = 0.7490

- Run 2 (development):
  - precision = 0.6149
  - recall = 0.8545
  - F = 0.7152

The results for the submitted runs were the following:

- Run 1 (test):
  - precision = 0.467
  - recall = 0.7168
  - F = 0.5655

- Run 2 (test):
  - precision = 0.4652
  - recall = 0.7079
  - F = 0.5615

It is remarkable the recall achieved by the system. The difference between precision and recall values is due to the high amount of YES predictions given by the system when there is not entailment in fact. This amount of YES predictions is quite similar to the amount of NO predictions given when there is not entailment. Then, the system must be improved in order to increase the amount of correct NO predictions.

In order to have an idea about the performance of this system with respect to other contemporary textual entailment recognizers, accuracy – evaluation measure of the RTE Challenge [1][2] – was calculated for it during the development time, showing values ranging 0.66 and 0.70. These values are in the state-of-the-art of this kind of systems [1].

5. Conclusion

This first experience with textual entailment recognizers for the Spanish language shows quite interesting results. Despite the developed system is not very complex, based on lexical analysis, its accuracy values suggest that it is well situated in the state-of-the-art of this kind of systems. A first question arises when comparing the present system with more complex ones previously developed by the authors for the English language [4] [5]: why so different systems achieve similar results? It is not easy to answer; the language, the kind of task and the different external resources used for the development are some factors that can affect the overall results.

As future work, this system should be improved in order to increase the amount of NO predictions given when there is not entailment in fact. For this, new features should be searched to study their contribution to the overall performance of the system and under what circumstances is interesting to consider or not every feature.

An associated task arises from the need of resources for the new languages involved in textual entailment tasks. Thus, it is necessary to develop and make available new tools such as, for example, dependency analyzers, with the quality of the ones existing for the English language.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Technology within the project: TIC-2003-07158-C04-02 Multilingual Answer Retrieval Systems and Evaluation, SyEMBRA, and a UNED PhD grant.

References

1. R. Bar-Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, Idan Szpektor. The Second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venezia, Italy, April 2006.
2. I. Dagan, O. Glickman, and B. Magnini. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop*, Southampton, UK. April 2005. LNAI, Springer.
3. S. Fischer, R. Klinkenberg, I. Mierswa, O. Ritthoff. Yale 3.0, Yet Another Learning Environment. User Guide, Operator Reference, Developer Tutorial. Technical Report. University of Dortmund, Department of Computer Science. Dortmund, Germany. 2005.
4. J. Herrera, A. Peñas, F. Verdejo. Textual Entailment Recognition Based on Dependency Analysis and WordNet. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop.* Southampton, UK. April 2005. LNAI, Springer.
5. J. Herrera, A. Peñas, Á. Rodrigo, F. Verdejo. UNED at PASCAL RTE-2 Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, pages 38-43, Venezia, Italy, April 2006.