UNED at PASCAL RTE-2 Challenge

Jesús Herrera, Anselmo Peñas, Álvaro Rodrigo, Felisa Verdejo
Departmento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
Madrid, Spain
{jesus.herrera, anselmo, alvarory, felisa}@lsi.uned.es

Abstract

This paper reports the description of the developed system and the results obtained in the participation of the UNED[1] in the Second Recognizing Textual Entailment (RTE) Challenge. New techniques and tools have been added: enriched queries to *WordNet*, detection of numeric expressions and their entailment, and Support Vector Machine classification (SVM) are the more relevant. The accuracy performed is slightly higher than the one from the previous edition system.

1 Introduction

The system presented to the Second Recognizing Textual Entailment Challenge is based on the one presented to the First RTE Challenge (Herrera et al., 2005). The core of this latter was basically kept, but enhanced by means of several subsystems in order to study the efficiency of other not previously applied techniques that seemed promising for RTE.

In short, the techniques involved in this new system are the following:

- Dependency analysis of texts and hypothesises.
- Lexical entailment between dependency tree nodes using *WordNet*. The subsystem consulting *WordNet* was enriched with respect to the one presented to the First RTE Challenge.
- Mapping between dependency trees, which is the one defined for the previous system (Herrera et al., 2005).
- Detection of numeric expressions. A new module, which detects entailment between numeric expressions of the texts and the hypothesises, was implemented. For this detection, the train and test corpora were automatically tagged (cardinals, dates and named entities) by the López-Ostenero's system (Peinado et al., 2005).
- Support Vector Machine classification in order to determine the final decision about textual entailment between pairs of text and hypothesis, following previous ideas from successful works in Natural Language Processing using machine learning applications (Joachims, 1998).

2 System Description

The proposed system is based on surface techniques of lexical and syntactic analysis, complemented with queries to *WordNet* as an external source of knowledge. It works in a non-specific way, not giving any kind of special treatment for the different settings considered in the RTE Challenge (Information Retrieval, Multi-document summarization, Question Answering and Information Extraction).

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 the learned model by a SVM classifier.

[1] Spanish Distance Learning University.

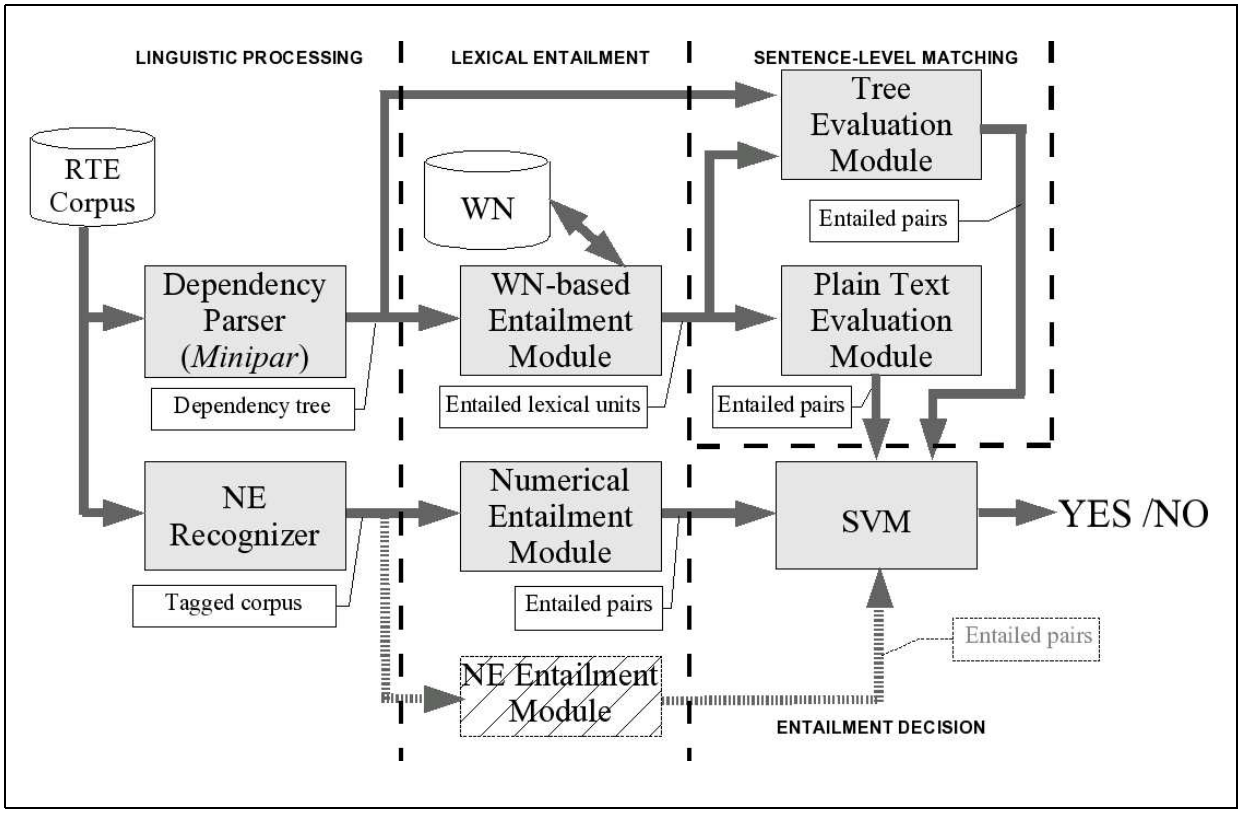

Figure 1: System's architecture.

System's components, whose graphic representation is shown in figure 1, are the following:

2.1 Linguistic processing

A dependency parser, based on Lin's Minipar (Lin, 1998), which normalizes data from the corpus of text and hypothesis pairs and accomplishes the dependency analysis, generating a dependency tree for every text and hypothesis.

A named entities recognizer, implemented by López-Ostenero (Peinado et al., 2005), has been used to normalize numeric expressions and named entities.

2.2 Lexical entailment

A *WordNet*-based entailment module – which takes the information given by the parser and returns the hypothesis' nodes that are entailed by the text (Herrera et al., 2005) – uses *WordNet* in order to find synonymy, similarity, hyponymy, *WordNet*'s entailment and negation relations between pairs of lexical units, as described in (Herrera et al., 2005). For the current edition some features have been added to the lexical entailment module:

- Search of entailment paths. It has been studied whether the strategy for searching the entailment paths affects or not the results. Two strategies have been tested: depth search and breadth search. In addition, the length of the path has been used as a final criteria for deciding the entailment between words. Although behaviour is slightly different, the effect over results in the exercise is no significant. However, the breadth strategy is significatively slower.

- *WordNet* relations. Synonymy, hyponymy, verb entailment and antonymy have been used as in the last edition. In addition, part meronym (e.g. *Italy* entails *Europe*), and adjective / adverb pertainym (e.g. *Italian* entails *Italy*) have been added in the search of entailment paths between the lemmas of the text and the ones of the hypothesis.

Table 1: Entailment between numeric expressions

	Text	Hypothesis
Recognition	*17 million citizens*	*more than 15 million people*
Normalization	lower bound: $17,000,000$ upper bound: $17,000,000$ unit: *citizen*	lower bound: $15,000,000$ upper bound: *infinite* unit: *person*
Entailment	*TRUE* if $15,000,000 \leq 17,000,000$ and *infinite* $\geq 17,000,000$ and *citizen* entails *person*	

- Entailment between phrases / multiwords. Levensthein distance has been used for an approximate matching between multiwords only if the one related to the hypothesis is present in *WordNet*. A new and simple entailment relation between phrases has been defined assuming the compositional meaning of phrases. Thus, a phrase is expected to entail all its components. This entailment relation can't be used over Named Entities since they haven't a compositional meaning.

- Entailment between numeric expressions. Numeric expressions from the corpus are detected by means of an entities recognizer; they are normalized after the recognition, and the units affected (e.g. kilometers, years, etcetera) are considered for the detection of an entailment relation between these expressions. Thus, a numeric expression N1 entails a numeric expression N2 if the range associated to N2 encloses the range of N1 and the unit of N1 entails the unit of N2. 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. An example is shown in table 1. The experiment in figure 2 shows the accuracy obtained over the development corpus when considering: coincidence between lemmas (*LEM*), *WordNet* relations (*WN*), and entailment between numeric expressions (*NUMFAIL*). For every percentage of overlap between the text and the hypothesis, the accuracy obtained is higher when adding *NUMFAIL* to the set of considered features; and the lower is the overlap the higher is this accuracy improvement. Thus, *NUMFAIL* is an interesting feature to decide if there is entailment between a text and a hypothesis when a lower overlap exists.

The named entities (NE) entailment module, as described in section 5, did not contribute to the runs submitted.

2.3 Sentence level matching

A tree matching module, which searches for matching branches into the hypothesises' dependency trees. These kind of branches are the ones whose all nodes are lexically entailed, as described in (Herrera et al., 2005).

A plain text matching module that calculates the percentage of lemmas from the hypothesis entailed by lemmas from the text, according to section 2.2.

2.4 Entailment decision

A SVM classifier, from Yet Another Learning Environment Yale 3.0 (Fischer et al., 2005), which 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 nodes of the hypothesis' dependency tree pertaining to matching branches (Herrera et al., 2005) considering, respectively:
   - Lexical entailment between the words of the snippets involved, without consulting *WordNet*.
   - Lexical entailment between the lemmas of the snippets involved, consulting *WordNet*.
2. Percentage of lemmas of the hypothesis entailed by lemmas of the text, considering the lexical entailment relations described in section 2.2.

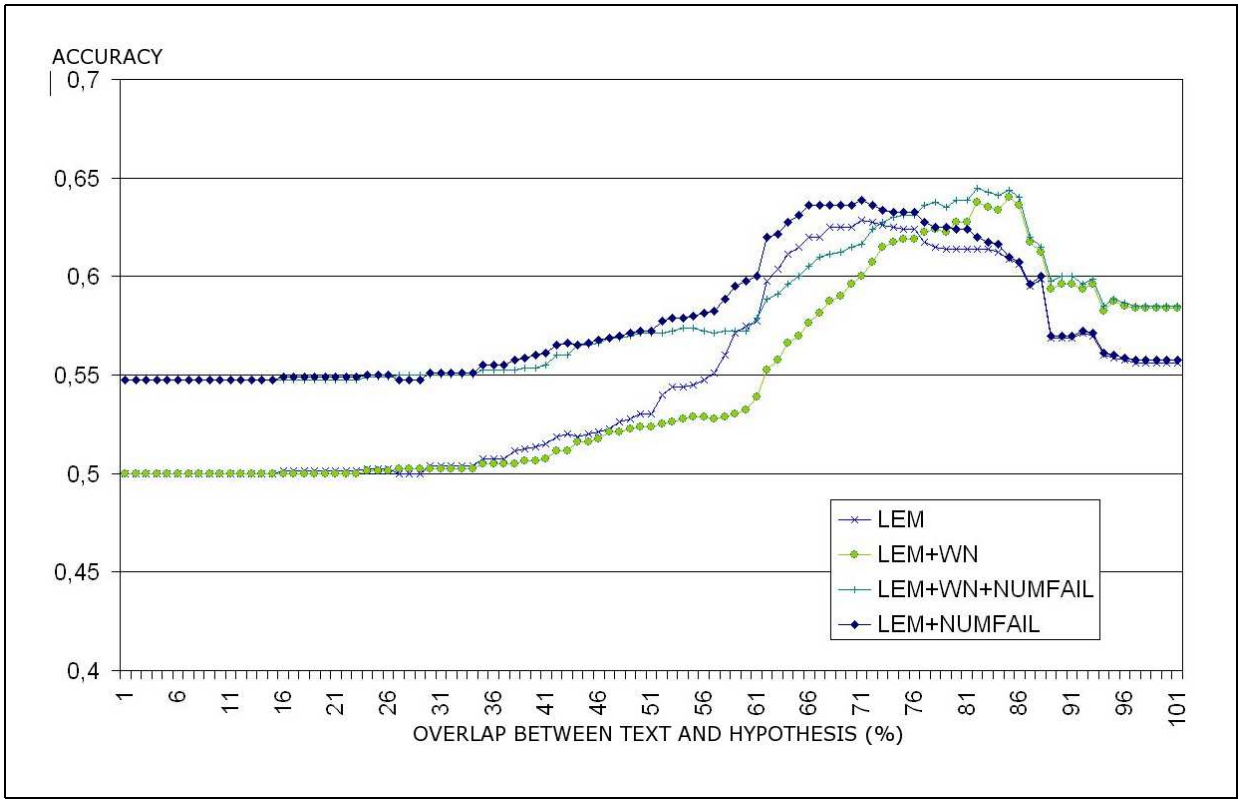

Figure 2: Effect of the numerical entailment restriction over the development corpus

3. Percentage of words of the hypothesis in the text (treated as bags of words).

4. Percentage of lemmas of the hypothesis in the text (treated as bags of lemmas).

5. Existence or absence of any numeric expression within the hypothesis.

6. Existence or absence of any numeric expression within the text.

7. Existence or absence of entailment between numeric expressions of the text and the hypothesis, as described in section 2.2.

3 Runs Submitted

Two runs were submitted to the Second RTE Challenge.

Run 1 was obtained using only the features 2 and 7 of section 2.4.

Run 2 was produced by the system described in section 2. The SVM was trained with the features enumerated in section 2.4, obtained from the development corpus provided by the organizers. Thus, the model was applied to the features from the test corpus in order to obtain a prediction for the existence or absence of textual entailment for every pair <text, hypothesis>.

4 Performance

Two evaluation measures were applied to the participating systems: accuracy (Dagan et al. , 2005), as the main measure, and average precision (Voorhees and Harman, 1999), as the secondary measure. Accuracy was computed for all the runs submitted, but average precision only for the runs giving the results ranked according to their entailment confidence; this ranking was not mandatory.

One of the two runs submitted was ranked and, then, average precision was computed for it. The results obtained over the test corpus are shown in tables 2 and 3.

The accuracy of the system is not homogeneous over all the pairs and depends on the different appli-

Table 2: Results for run 1

	Accuracy	Average Precision
IE	49.00%	47.74%
IR	64.50%	69.14%
QA	56.50%	50.24%
SUM	69.00%	79.05%
Overall	59.75%	56.63%

Table 3: Results for run 2

	Accuracy
IE	52.00%
IR	57.00%
QA	52.00%
SUM	74.50%
Overall	58.87%

cation settings proposed by the organizers: Information Retrieval (IR), Multi-document Summarization (SUM), Question Answering (QA) and Information Extraction (IE). The overall accuracy shown by the current system is basically due to the contribution of the Multi-document Summarization setting. This setting is characterized by sentence pairs with high lexical overlap, and the system shows its better accuracy for the subset of pairs pertaining to this kind of setting, for which reaches 74.50% accuracy.

4.1 Performance comparison

Though the overall accuracy is better than the one obtained in the First RTE Challenge, the improvement has not been very significant: only 3.38 percentage points between the best performances reached in every edition. Considering a combination of the two systems used to obtain the runs submitted to de Second RTE Challenge, in which the IR and QA pairs were treated by the system producing run 1 and the IE and SUM pairs were treated by the system producing run 2, the best performance could be given. In such a case, the overall accuracy will be 61.88%; thus, the performance improvement with respect to the previous edition of the Challenge will be of 5.51 percentage points.

During the development of the system, some experiments were accomplished in order to compute its accuracy. Training the SVM with a half of the development corpus and applying the model to the other half, and vice-versa, the overall accuracy obtained ranged between 63% and 64%. It overcomes in more than 4 percentage points the best accuracy obtained by the two runs submitted. Since the development corpus and the test corpus are similar, it can be concluded that the accuracy of the system is quite variable depending on the concrete samples of the corpus. Then, it is not easy to affirm that the current system is clearly better than the previous one.

In the first edition of the RTE Challenge the Multi-document Summarization setting was not proposed but a similar one called Comparable Documents (Dagan et al. , 2005), characterized by a high lexical overlap between texts and hypothesises, too. For Comparable Documents, the system presented to the previous edition of the RTE Challenge reached its best accuracy, with a 79.33%. The accuracy obtained for the settings with a high lexical overlap went slightly down from the first to the second edition. Because of the settings are not exactly the same, no definitely conclusions can be stated, but it is clear that both systems are significantly good at recognizing textual entailment when the pairs show a high lexical overlap.

5 New Experiments after Second RTE Challenge Participation

After submitting the results to the Second RTE Challenge, the development of the system continued and some experiments were accomplished.

5.1 New features based on named entities

A new module for recognizing entailment between named entities (see figure 1) was implemented in order to study its usefulness for RTE. It works in a similar way to the numerical entailment module. The features computed were the following: a) existence or absence of any named entity within the hypothesis; b) existence or absence of any named entity within the text; c) existence or absence of entailment between named entities of the text and the hypothesis; it is said that there is entailment between named entities of the text and its correspondant hypothesis if, in case of existence of named entities in the hypothesis, all them are entailed by one ore more named entities from the text.

The results obtained after executing the whole system described in figure 1 – over the test corpus – are shown in table 4.

Table 4: Results considering named entities

	Accuracy
IE	51.00%
IR	63.50%
QA	53.50%
SUM	74.00%
Overall	60.50%

The overall accuracy of this system is slightly better than the ones obtained by the other two systems. Thus, named entities seem to be an interesting field of study in order to improve RTE.

6 Conclusions and Future Work

The system presented to the First RTE Challenge has been improved in order to put in the same level the weight of the analysis due to the overlap between dependency trees – which represented the only decision item in the previous system – and the weight of other kinds of analysis, such as bag of words, bag of lemmas, numerical entailment, etcetera. The relevance of every feature computed for the pairs <text, hypothesis> has been determined automatically by a SVM algorithm (training with the development corpus) which is the responsible for the prediction of existence/absence of entailment between the pairs of the test corpus.

Despite the complexity of the developed system is quite higher than the complexity of the previous one, the obtained accuracy is slightly better. It seems not easy to determine the way to obtain a higher accuracy for every application setting. With the currently experimented techniques and tools, the results obtained for settings with a high lexical overlap are significantly higher than the others, which are quite similar among themselves. It suggest that it should be stimulated the development of setting-oriented systems, aiming to increase the performance of RTE systems focusing on only one or a few settings. Thus, in the medium term, useful RTE-based systems for specific uses could be hopefully available.

From the results obtained along the development time and after the execution of the test it can be deduced that, nowadays, the RTE systems could be used with a remarkable success to identify information redundancy in multi-document summarization tasks.

Other applications for RTE systems should be explored, such as automatic Answer Validation tasks. An example for this kind of task is proposed within the Cross Language Evaluation Forum (CLEF) for the year 2006[2].

7 Acknowledgments

We are grateful to Fernando López-Ostenero, from UNED-NLP Group, for his named entities recognizer.

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.

References

I. Dagan, O. Glickman and B. Magnini. 2005. *The PASCAL Recognising Textual Entailment Challenge*. Proceedings of the First PASCAL Recognizing Textual Entailment Workshop. LNAI. Springer. In press.

S. Fischer, R. Klinkenberg, I. Mierswa and O. Ritthoff. 2005. *Yale 3.0, Yet Another Learning Environment. User Guide, Operator Reference, Developer Tutorial*. University of Dortmund, Department of Computer Science. Dortmund, Germany.

J. Herrera, A. Peñas and F. Verdejo. 2005. *Textual Entailment Recognition Based on Dependency Analysis and WordNet*. Proceedings of the First PASCAL Recognizing Textual Entailment Workshop. LNAI. Springer. In press.

T. Joachims. 1998. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. LS-8 Report 23. University of Dortmund, Department of Computer Science. Dortmund, Germany.

D. Lin. 1998. *Dependency-based Evaluation of MINIPAR*. Workshop on the Evaluation of Parsing Systems, Granada, Spain, May, 1998.

V. Peinado, F. López-Ostenero, J. Gonzalo and F. Verdejo. 2005. *UNED at ImageCLEF 2005: Automatically Structured Queries with Named Entities over Metadata*. Cross Language Evaluation Forum, Working Notes for the CLEF 2005 Workshop. LNCS. Springer. In press.

M. Voorhees and D. Harman. 1999. *Overview of the seventh text retrieval conference*. Proceedings of the Seventh Text Retrieval Conference (TREC-7). NIST Special Publication.

[2] http://nlp.uned.es/QA/AVE/