Machine Learning with Semantic-Based Distances Between Sentences for Textual Entailment

Daniel Ferrés
TALP Research Center
Software Department
Universitat Politècnica de Catalunya
dferres@lsi.upc.edu

Horacio Rodríguez
TALP Research Center
Software Department
Universitat Politècnica de Catalunya
horacio@lsi.upc.edu

Abstract

This paper describes our experiments on Textual Entailment in the context of the Third Pascal Recognising Textual Entailment (RTE-3) Evaluation Challenge. Our system uses a Machine Learning approach with Support Vector Machines and AdaBoost to deal with the RTE challenge. We perform a lexical, syntactic, and semantic analysis of the entailment pairs . From this information we compute a set of semantic-based distances between sentences. The results look promising specially for the QA entailment task.

1 Introduction

This paper describes our participation in the RTE-3 challenge. It is our first attempt to RTE and we have taken profit of an analysis of the approaches followed in previous challenges (see (Dagan et al., 2005), and (Bar-Haim et al., 2006) for overviews of RTE-1 and RTE-2). Our approach, however, is based on a set of semantic-based distance measures between sentences used by our group in previous contests in Question Answering (TREC 2004, see (Ferrés et al., 2005), and CLEF 2004, see (Ferrés et al., 2004)) , and Automatic Summarization (DUC 2006, see (Fuentes et al., 2006)). Although the use of such measures (distance between question and sentences in passages candidates to contain the answer, distance between query and sentences candidates to be included in the summary, ...) is different for RTE task, our claim is that with some modifications the approach can be useful in this new scenario.

The organization of this paper is as follows. After this introduction we present in section 2 a description of the measures upon which our approach is built. Section 3 describes in detail our proposal. Results are discussed in section 4. Conclusions and further work is finally included in section 5.

2 System Description

Our approach for computing distance measures between sentences is based on the degree of overlapping between the semantic content of the two sentences. Obtaining the semantic content implies a depth Linguistic Processing. Upon this semantic representation of the sentences several distance measures are computed. We next describe such issues.

2.1 Linguistic Processing

Linguistic Processing (LP) consists of a pipe of general purpose Natural Language (NL) processors that performs tokenization, morphologic tagging, lemmatization, Named Entities Recognition and Classification (NERC) with 4 basic classes (PERSON, LOCATION, ORGANIZATION, and OTHERS), syntactic parsing and semantic labelling, with WordNet synsets, Magnini's domain markers and EuroWordNet Top Concept Ontology labels. The *Spear*[1] parser performs full parsing and robust detection of verbal predicate arguments. The syntactic constituent structure of each sentence (including the specification of the head of each constituent) and the relations among constituents (subject, direct and indirect object, modifiers) are obtained. As a result

[1]**Spear.** http://www.lsi.upc.edu/~surdeanu/spear.html

Proceedings of the Workshop on Textual Entailment and Paraphrasing, pages 60–65,
Prague, June 2007.

of the performance of these processors each sentence is enriched with a lexical and syntactic language dependent representations. A semantic language independent representation of the sentence (called *environment*) is obtained from these analyses (see (Ferrés et al., 2005) for details). The *environment* is a semantic network like representation built using a process to extract the semantic units (nodes) and the semantic relations (edges) that hold between the different tokens in the sentence. These units and relations belong to an ontology of about 100 semantic classes (as person, city, action, magnitude, etc.) and 25 relations (mostly binary) between them (e.g. *time_of_event*, *actor_of_action*, *location_of_event*, etc.). Both classes and relations are related by taxonomic links (see (Ferrés et al., 2005) for details) allowing inheritance. Consider, for instance, the sentence "Romano_Prodi 1 is 2 the 3 prime 4 minister 5 of 6 Italy 7". The following environment is built:

i_en_proper_person(1), *entity_has_quality(2)*, *entity(5)*, *i_en_country(7)*, *quality(4)*, *which_entity(2,1)*, *which_quality(2,5)*, *mod(5,7)*, *mod(5,4)*.

2.2 Semantic-Based Distance Measures

We transform each environment into a labelled directed graph representation with nodes assigned to positions in the sentence, labelled with the corresponding token, and edges to predicates (a dummy node, 0, is used for representing unary predicates). Only unary (e.g. *entity(5)* in Figure 1) and binary (e.g. in Figure 2 *which_quality(2,5)*) predicates are used. Over this representation a rich variety of lexico-semantic proximity measures between sentences have been built. Each measure combines two components:

- A lexical component that considers the set of common tokens occurring in both sentences. The size of this set and the strength of the compatibility links between its members are used for defining the measure. A flexible way of measuring token-level compatibility has been set ranging from word-form identity, lemma identity, overlapping of WordNet synsets, approximate string matching between Named Entities etc. For instance, "Romano Prodi" is lexically compatible with "R. Prodi" with a score of 0.5 and with "Prodi" with a score of 0.41. "Italy" and "Italian" are also compatible with score 0.7. This component defines a set of (partial) weighted mapping between the tokens of the two sentences that will be used as anchors in the next component.

- A semantic component computed over the subgraphs corresponding to the set of lexically compatible nodes (anchors). Four different measures have been defined:
  - Strict overlapping of unary predicates.
  - Strict overlapping of binary predicates.
  - Loose overlapping of unary predicates.
  - Loose overlapping of binary predicates.

The loose versions allow a relaxed matching of predicates by climbing up in the ontology of predicates (e.g. provided that A and B are lexically compatible, *i_en_city(A)* can match *i_en_proper_place(B)*, *i_en_proper_named_entity(B)*, *location(B)* or *entity(B)*) [2]. Obviously, loose overlapping implies a penalty on the score that depends on the length of the path between the two predicates and their informative content.

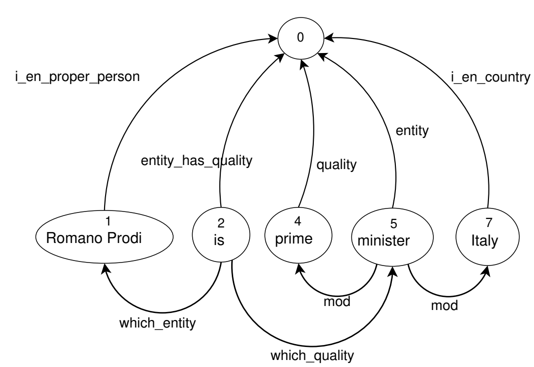

Figure 1: Example of an environment of a sentence.

[2]The ontology contains relations as *i_en_city* **isa** *i_en_proper_place*, *i_en_proper_place* **isa** *i_en_proper_named_entity*, *proper_place* **isa** *location*, *i_en_proper_named_entity* **isa** *entity*, *location* **isa** *entity*

3 System Architecture

We have adapted the set of measures described before for RTE in the following way:

1. We follow a Machine Learning (ML) approach for building a classifier to perform the RTE task. In previous applications the way of weighting and combining the different measures was based on a crude optimization using a development corpus.

2. We extract a more complex set of features for describing the semantic content of the Text (T) and the Hypothesis (H) as well as the set of semantic measures between them. Table 1 contains a brief summary of the features used.

3. We perform minor modifications on the token-level compatibility measures for dealing with the asymmetry of the entailment relation (basically using the hyponymy and the verbal entailment relations of WordNet)

4. We add three new task-specific features (see Table 1)

The overall architecture of the system is depicted in Figure 2. As usual in ML, the system proceeds in two phases, learning and classification. The left side of the figure shows the learning process and the right part the classification process. The set of examples (tuples H, T) is first processed, in both phases, by LP for obtaining a semantic representation of the tuple (H_{sem} and T_{sem}). From this representation a Feature Extraction component extracts a set of features. This set is used in the learning phase for getting a classifier that is applied to the set of features of the test, during the classification phase, in order to obtain the answer.

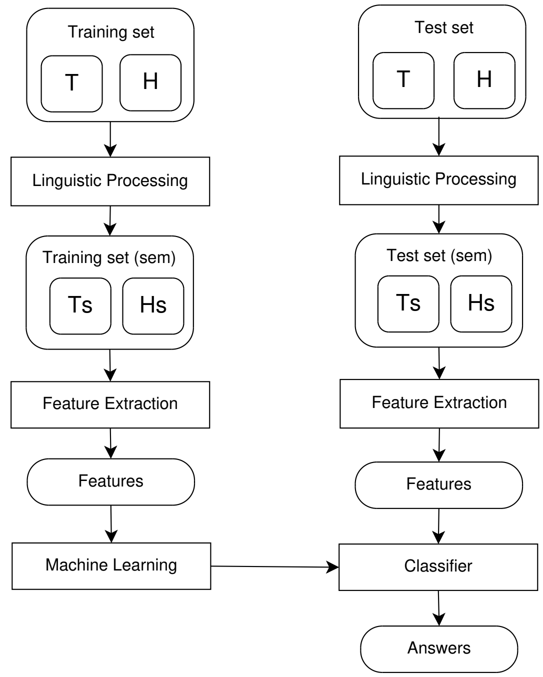

Figure 2: System Architecture.

4 Experiments

Before the submission we have performed a set of experiments in order to choose the Machine Learning algorithms and the training sets to apply in the final submission.

4.1 Machine Learning Experiments

We used the WEKA[3] ML platform (Witten and Frank, 2005) to perform the experiments. We tested 9 different ML algorithms: *AdaBoostM1*, *Bayes Networks*, *Logistic Regression*, *MultiBoostAB*, *Naive Bayes*, *RBF Network*, *LogitBoost* (Simple Logistic in WEKA), *Support Vector Machines* (SMO in WEKA), and *Voted Perceptron*. We used the previous corpora of the RTE Challenge (RTE-1 and RTE-2) and the RTE-3 development test. A filtering process has been applied removing pairs with more than two sentences in the text or hypothesis, resulting a total of 3335 Textual Entailment (TE) pairs. The results over 10-fold-Cross-Validation using a data set composed by RTE-1, RTE-2, and RTE-3 development set are shown in Table 2.

The results shown that *AdaBoost*, *LogitBoost*, and *SVM* obtain the best results. Then we selected *AdaBoost* and *SVM* to perform the classification of the RTE-3 test set. The *SVM* algorithm tries to compute the hyperplane that best separates the set of training examples (the hyperplane with maximum margin) (Vapnik, 1995). On the other hand, *AdaBoost* com-

[3]**WEKA.** `http://www.cs.waikato.ac.nz/ml/weka/`

Features	#features	Description
semantic content of T	12	#locations, #persons, #dates, #actions, ...
semantic content of H	12	...
intersection of T and H	12	...
Strict overlapping of unary predicates	5	length of intersection score of intersection ratio of intersection related to shortest env ratio of intersection related to longest env ratio of intersection related to both (union of)
Strict overlapping of binary predicates	5	...
Loose overlapping of unary predicates	5	...
Loose overlapping of binary predicates	5	...
Verbal entailment (WordNet)	1	V1 ϵ T, V2 ϵ H, such that V1 verbal_entails V2
Antonymy	1	A1 ϵ T, A2 ϵ H, such that A1 and A2 are antonyms and no token compatible with A2
#occurs in H Negation	1	Difference between # negation tokens in H and T

Table 1: Features used for classification with Machine Learning algorithms.

Algorithm	#correct	Accuracy
AdaBoostM1	1989	59.6402
BayesNet	1895	56.8216
Logistic	1951	58.5007
MultiBoostAB	1959	58.7406
NaiveBayes	1911	57.3013
RBFNetwork	1853	55.5622
LogitBoost	1972	59.1304
SVM	1972	59.1304
VotedPerceptron	1969	59.0405

Table 2: Results over 10-fold-Cross-Validation using a filtered data set composed by RTE-1, RTE-2, and RTE-3 (a total of 3335 entailment pairs).

bines a set of weak classifiers into a strong one using lineal combination (Freund and Schapire, 1996). The idea is combining many moderately accurate rules into a highly accurate prediction rule. A weak learning algorithm is used to find the weak rules.

4.2 Training Set Experiments

We designed two experiments in order to decide the best training set to apply in the RTE-3 challenge. We performed an experiment using RTE-1 and RTE-2 data sets as a training set and the RTE-3 development set filtered (541 TE pairs) as a test set. In this experiment *AdaBoost* and *SVM* obtained accuracies of 0.6672 and 0.6396 respectively (see results in Table 3. We performed the same experiment joining the Answer Validation Exercise[4] (AVE) 2006 English data set (Peñas et al., 2006) and the Microsoft Research Paraphrase Corpus[5] (MSRPC) (Dolan et al., 2004) to the previous corpora (RTE-1 and RTE-2) resulting a total of 8585 entailment pairs filtering pairs with a text or a hypothesis with more than 1 sentence. In our approach we considered that paraphrases were bidirectional entailments. The paraphrases of the MSRPC have been used as textual entailments in only one direction: the first sentence in the paraphrase has been considered the hypothesis and the second one has been considered the text.

Using the second corpus for training and the RTE-3 development set as test set resulted in a notable degradation of accuracy (see Table 3).

	Accuracy	
Algorithm	Corpus A	Corpus B
AdaBoost	66.72%	53.78%
SVM	63.95%	59.88%

Table 3: Results over the RTE-3 development set filtered (541 TE pairs) using as training set corpus A (RTE-1 and RTE-2) and corpus B (RTE-1, RTE-2, MSRPC, and AVE2006 English)

Finally, we performed a set of experiments to detect the contribution of the different features used for Machine Learning. These experiments revealed that

[4]**AVE.** http://nlp.uned.es/QA/AVE

[5]**MSRPC.** http://research.microsoft.com/nlp/msr_paraphrase.htm

the three most relevant features were: Strict overlapping of unary predicates, Semantic content of Hypothesis, and Loose overlapping of unary predicates.

4.3 Official Results

Our official results at RTE-3 Challenge are shown in Table 4. We submitted two experiments: the first one with *AdaBoost* (run1) and the second one with *SVM* (run2). Training data set for final experiments were corpus: RTE-1 (development and test), RTE-2 (development and test), and RTE-3 development. The test set was the RTE-3 test set without filtering the entailments (text or hypothesis) with more than one sentence. In this case we joined multiple sentences in a unique sentence that has been processed by the LP component.

We obtained accuracies of 0.6062 and 0.6150. In the QA task we obtained the best per-task results with accuracies of 0.7450 and 0.7000 with *AdaBoost* and *SVM* respectively.

	Accuracy	
Task	run1 *AdaBoost*	run2 *SVM*
IE	0.4350	0.4950
IR	0.6950	0.6800
QA	0.7450	0.7000
SUM	0.5500	0.5850
Overall	0.6062	0.6150

Table 4: RTE-3 official results.

5 Conclusions and Further Work

This paper describes our experiments on Textual Entailment in the context of the Third Pascal Recognising Textual Entailment (RTE-3) Evaluation Challenge. Our approach uses Machine Learning algorithms (*SVM* and *AdaBoost*) with semantic-based distance measures between sentences. Although further analysis of the results is in process, we observed that our official per-task results at RTE-3 show a different distribution compared with the global results of all system at RTE-2 challenge. The RTE-2 per-task analysis showed that most of the systems scored higher in accuracy in the multidocument summarization (SUM) task while in our system this measure is low. Our system at RTE-3 challenge scored higher in the QA and IR tasks with accuracies of 0.7450 and 0.6950 respectively in the first run.

References

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment.*

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In Joaquin Quiñonero Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *MLCW*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 350, Morristown, NJ, USA. Association for Computational Linguistics.

Daniel Ferrés, Samir Kanaan, Alicia Ageno, Edgar González, Horacio Rodríguez, Mihai Surdeanu, and Jordi Turmo. 2004. The TALP-QA System for Spanish at CLEF 2004: Structural and Hierarchical Relaxing of Semantic Constraints. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors, *CLEF*, volume 3491 of *Lecture Notes in Computer Science*, pages 557–568. Springer.

Daniel Ferrés, Samir Kanaan, Edgar González, Alicia Ageno, Horacio Rodríguez, Mihai Surdeanu, and Jordi Turmo. 2005. TALP-QA System at TREC 2004: Structural and Hierarchical Relaxation Over Semantic Constraints. In *Proceedings of the Text Retrieval Conference (TREC-2004).*

Yoav Freund and Robert E. Schapire. 1996. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pages 148–156.

Maria Fuentes, Horacio Rodríguez, Jordi Turmo, and Daniel Ferrés. 2006. Femsum at duc 2006: Semantic-based approach integrated in a flexible eclectic multitask summarizer architecture. In *Proceedings of the Document Understanding Conference 2006 (DUC 2006). HLT-NAACL 2006 Workshop.*, New York City, NY, USA, June.

Anselmo Peñas, Álvaro Rodrigo, Valentín Sama, and Felisa Verdejo. 2006. Overview of the answer validation exercise 2006. In *Working Notes for the*

CLEF 2006 Workshop. ISBN: 2-912335-23-x, Alicante, Spain, September.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, June.

Acknowledgments

This work has been supported by the Spanish Research Dept. (TEXT-MESS, TIN2006-15265-C06-05). Daniel Ferrés is supported by a UPC-Recerca grant from Universitat Politècnica de Catalunya (UPC). TALP Research Center is recognized as a Quality Research Group (2001 SGR 00254) by DURSI, the Research Department of the Catalan Government.