TALP at TAC 2008: A Semantic Approach to Recognizing Textual Entailment

Alicia Ageno*, **David Farwell*** , **Daniel Ferrés*** , **Fermín Cruz**† , **Horacio Rodríguez*** , and **Jordi Turmo***

*TALP Research Center
Universitat Politècnica de Catalunya
Spain
{dferres,horacio}@lsi.upc.edu

† Universidad de Sevilla
Spain
fermincruz@gmail.com

Abstract

This paper describes our experiments on Textual Entailment in the context of the Fourth Recognising Textual Entailment (RTE-4) Evaluation Challenge at TAC 2008 contest. Our system uses a Machine Learning approach with 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. We improved our baseline system for the RTE-3 challenge with more Language Processing techniques, an hypothesis classifier, and new semantic features. The results show no general improvement with respect to the baseline.

1 Introduction

This paper describes our participation in the RTE-4 challenge within TAC 2008 contest. It is our second participation in RTE exercises, after our participation, last year, in the RTE-3 challenge, within Pascal program, (Ferrés and Rodríguez, 2007).

In previous challenges, a large number of techniques have been applied to the task (see (Dagan et al., 2005), (Bar-Haim et al., 2006), and (Giampiccolo et al., 2007) for overviews of RTE-1, RTE-2, and RTE-3).

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 core of our system this year follows basically the same approach of our previous system but differs in three aspects from it:

1. We have improved some of the linguistic components of the system and incorporate several additional lexical resources.

2. We have added a new component, an hypothesis classifier. The aim of this component is to group together the hypothesis into coherent classes for allowing the computation of finer features to be included in the RTE classifier.

3. The set of features of the RTE classifier has been notably increased.

The organization of this paper is as follows. After this introduction we present in section 2 a brief description of our basic system, focusing on the measures upon which our approach is built (for a more detailed description, see (Ferrés and Rodríguez, 2007)). Section 3 describes in detail the improvements we have included this year. Results are shown in section 4. Conclusions and further work is finally included in section 5.

2 Basic System Description

The overall architecture of the system is depicted in Figure 1. 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 a Linguistic Processing (LP) component for obtaining a semantic representation of the tuple (H_{sem} and T_{sem}). From this representation a Feature Extraction (FE) 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, for producing the answer.

Most features are computed from the comparison between H_{sem} and T_{sem}. Our approach for computing distance measures between sentences is based on the degree of overlapping between the semantic content of the two sentences. Obtaining this semantic content implies an in depth Linguistic Processing (LP) described in section 2.1. Upon this semantic representation of the sentences several distance measures are computed, as described in section 2.2. These measures form the base of the set of features used by our RTE classifier, as described in section 2.3.

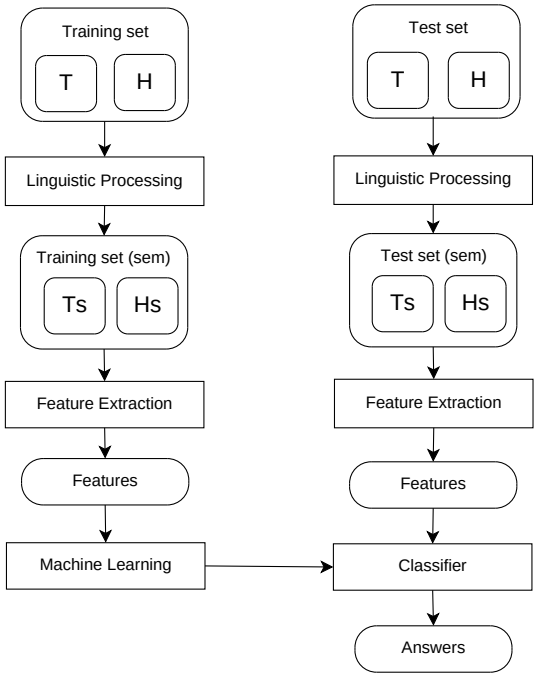

Figure 1: System Architecture.

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 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 corre-

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

sponding token, and edges to predicates (a dummy node, 0, is used for representing unary predicates). Only unary (e.g. *entity(5)* in Figure 2) 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 strict versions imply that two predicates exactly match being their arguments lexically compatible. 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.

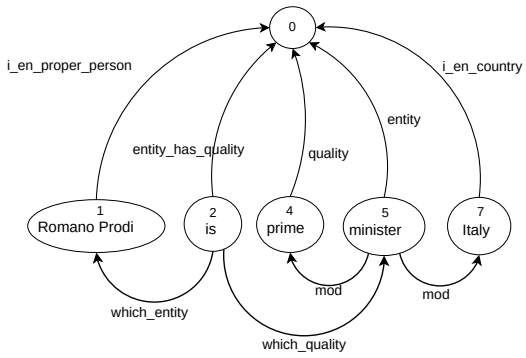

Figure 2: Example of an environment of a sentence.

2.3 RTE classifier

Last year we used the WEKA[3] ML platform (Witten and Frank, 2005) to perform some experiments for choosing the most appropriate classifier. 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 was applied removing pairs with more than two sentences in the text or hypothesis, resulting a total of 3,335 Textual Entailment (TE) pairs. The results shown that AdaBoost, LogitBoost, and SVM obtain the best results. The difference of performance between the different classifiers were small and not statistically significant for the first three learners (as the feature set was the same for all the systems this result is not surprising). Then we selected AdaBoost and SVM to perform the classification of the RTE-3 test set. Our two runs in RTE-3 used respectively these two learners. This year we used only AdaBoost and we have included multi-sentence pairs for learning.

Last year we experimented with different data sets for learning:

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

1. Using only the training material provided by the organizers, i.e. RTE-1 and RTE-2 both training and test and RTE-3 only training.

2. Using complementary material, namely 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). We performed the same experiment joining MSRPC, and AVE 2006 English.

As the inclusion of complementary learning material resulted in a fall in accuracy this year we have used for learning only the datasets (training and test) of previous challenges.

In our participation last year we used a set of features based on the semantic overlap between the Text (T) and the Hypothesis (H) . Table 1 contains a brief summary of the features used:

First we used 12 features for summarizing the semantic content of T (and the corresponding of H). These features correspond to frequencies of locations, persons, organizations, dates, times, magnitudes, units of measure, actions, etc. The same features are used for describing the intersection (in terms of predicates) of T and H. In this way we got the 36 first features in Table 1. We added 20 features measuring the degree of overlapping of the predicates in T and H (based on the use of compatible predicate over all the possible mappings between the terms in T and H) . We used different features for unary and binary predicates and for strict and loose overlapping, as described in section 2.2. We include absolute and relative measures of this overlapping and also a score computed by combining the scores assigned by the compatible predicate.

Note that all these features capture simply the degree of overlapping between T and H, they are appropriate, thus, for detecting paraphrases. For proper entailments we added a new predicate entails with the same signature as compatible. In our previous system this predicate was reduced to use the corresponding relation in WN between verbs. Finally we added two features for dealing with antonyny, also using WN as Knowledge Source, and negation.

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

[5]**MSRPC**. `http://research.microsoft.com/`

3 Improvements on the basic system

As we said in the introduction, our current system incorpores several changes in three directions: improvements on the LP component, addition of hypothesis classifier and extension of the features set. We will in turn address these issues:

3.1 Improvements on the LP components

.

An analysis of the sources of error in our previous system revealed the following problems related to the LP components:

1. The lack of a coreference component supposed a severe drawback, specially in the case of multi-sentence texts.

2. The accuracy of our NERC component was poor. The system suffered a fall of about 10 points in accuracy with respect its usual performance on agency news and newspaper articles, probably due to the small context for disambiguating and the varieties of genres present in the corpus.

3. The compatible predicate failed to recognize many correct mappings, specially in the case of synonymy not covered in WN, approximate string matching of NEs, related words having different POS, etc.

4. The entails predicate coverage, reduced to verbs having the entailment relation in WN was clearly insufficient.

For facing these problems we performed the following actions:

1. We included in the LP pipe a simple coreference solver, reduced to recover pronominal co-references[6].

2. We developed a resegmentation/reclassification component. The aim is to resegment and/or to reclassify if needed the NEs proposed by our NER component when there is additional evidence supporting a new segmentation or classification.

[6]The co-reference solver is an in-house implementation. Thanks to Edgar González and Pere Comas for providing it.

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.

3. We enriched the set of compatible predicates using additional resources as WN relations and VerbOcean.

4. We extended the set of entails predicates.

Resegmentation can consist on: i) extending the original NE beyond its limits, i.e. incorporating one or two tokens before of after it, ii) merging two or more contiguous NEs into a new one, iii) splitting a NE into several ones, and iv) reducing the size of the NE taking away a prefix or a suffix. Additionally this component is able to reclassify the NE, i.e. changing its label or deciding that this is not a true NE. For this task we need additional resources. We have used the taxonomic information provided by two gazetteers: GNIS for USA toponyms, and Geonames for toponyms away from USA. Besides these resources we have used WN relations for toponyms included in WN, frequencies of the NE capitalized or not in a large corpus (the BNC) and categories attached to the NE in the Wikipedia. A severe constraint we impose is that all the occurrences of a NE in both T and H of a pair have the same label. As a result about 15% of the NEs have been changed (we set a rather restrictive threshold for allowing a change) with an accuracy of about 90%.

In our previous system, the available forms of compatibility between tokens provided by the compatible predicates were rather simple: word-form identity, lemma identity, overlapping of WordNet synsets, approximate string matching between Named Entities etc. We have extended the set including relations between actions and actors (as "work", "worker") and between locations and inhabitants (as "Spain", "Spanish"), relations between locations including or not trigger words (as "New York City" and "New York "), different forms of naming people (as "President Bush", "Bush", "Georges Bush", "Mr. Bush"), using of WN relations as "also see", using of Verbocean relations as "similar", different forms of acronym expansion, etc.

In our previous system, the available forms of entailment relations between tokens provided by the entails predicates were reduced to entailment between verbs set by the corresponding WN relation. We have extended the set including the use of other relations from WN as meronymy (part meronym, member meronym, substance meronym) and holonymy, as in "I visited Madrid" entails "I visited Spain" but not the inverse, the same kind of relations can be extracted from GNIS and Geonames gazetteers. Also the VerbOcean relations "stronger-than" and "happens-before" have been used.

We have added too entails predicates for managing dates (as "May 15th" entails "May"). Other more complex examples we are able to solve are " University_of_Milan" entails "Milan" or "Chicago_Blackhawks" entails "Chicago".

3.2 Classifying Hypothesis

In section 3.1 some improvements of our LP components have been presented, our aim was simply to improve the system but no novelty is proposed. Our second improvement on the previous system con-

sists of classifying the hypothesis into a set of possible classes and then applying finer measures of overlapping. This classifier is a clear novelty over our previous system. As most of the H include a single predicate, event in general, action and state in more specific cases, the pairs have been classified into the following classes (using only H information):

'aoo', 'apo', app', 'aco', 'aop', 'aoc', 'ape', 'aep', 'aoe', 'aeo', 'spp', 'sll', 'spo', 'epp', 'epo', 'epl', 'eop', 'eoo', 'eol', 'ell', 'ap*', 'ao*', 'ae*', 'sp*', 'so*', 'se*', 'a**', 's**', 'e**', '***'

where the first position refers to the predicate (a= action, s=state, e=event), the second to the subject and the third to the object (o= organization, p=person, l=location, e=whatever entity). * means that the position is not covered (or related). Class '***' is included as default value. Note that the different classes form a taxonomy being '***' the most general class.

The classification task is quite straightforward, using simple syntactic information, i.e. looking for the head of the predicates occurring in H and for their arguments.

3.3 New Features

We have added 30 new features corresponding to the presence and degree of satisfaction of some of the classes described in section 3.2. If an example does not belong to a class the value assigned to the feature is zero. In other case the value is computed multiplying the scores assigned by the compatible or entails predicates applied to the head and the two main arguments of the involved clause. For instance, if H contains an action with a person playing the role of subject and an organization playing the role of direct object, H has been classified as "apo" (and also as "ap*', 'a**', 'e**' and '***'). So all these 5 features have to be filled. In the case of "apo", the three involved tokens in H have to be mapped to three tokens in T, being each of the mappings related by compatible or entails predicates. The product of the corresponding scores gives to value of the feature. Obviously if one of the mapping does not occur the value results zero.

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. We choosed AdaBoost as a Machine Learning system and the set of all the RTE challenges as a training set.

4.1 Official Results

Our official results at RTE-4 Challenge 2-way task are shown in Table 2. We submitted three experiments with Adaboost: i) a baseline with the same features used at RTE-3 but an improved Linguistic Processing phase with coreference and NERC resegmentation/reclassification (run1), ii) baseline with added 30 new features based on presence and degree of satisfaction of the hypothesis classifier classes (run2). iii) (run3) baseline with the new 30 features in both training and test set but an enriched set of compatible and entails predicates using WordNet relations and VerbOcean has been applied in the test set.

The training data set for final experiments were the development and test sets of the RTE-1, RTe-2 and RTE-3. We obtained accuracies of 0.5630, 0.5540 and 0.5610 respectively.

Run	Description	Accuracy
run1	Baseline (RTE-3 features)	0.5630
run2	+30 features	0.5540
run3	+30 features + new predicates	0.5610

Table 2: RTE-4 official results.

5 Conclusions and Further Work

This paper describes our experiments on Textual Entailment in the context of the Fourth Recognising Textual Entailment (RTE-4) Evaluation Challenge. Our approach uses *AdaBoost* with semantic-based distance measures between sentences. We improved our baseline system for the RTE-3 challenge with more Language Processing techniques, an hypothesiss classifier, and new semantic features. The results show no general improvement with respect to the baseline. Further analysis of the results is in process.

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 and Horacio Rodríguez. 2007. Machine learning with semantic-based distances between sentences for textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 60–65, Prague, June. 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).*

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.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague, June. Association for Computational Linguistics.

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.

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.