An Approach to Cross-Lingual Textual Entailment using Online Machine Translation Systems

Julio Castillo and Marina Cardenas

Abstract—**In this paper, we show an approach to cross-lingual textual entailment (CLTE) by using machine translation systems such as Bing Translator and Google Translate. We experiment with a wide variety of data sets to the task of textual Entailment (TE) and evaluate the contribution of an algorithm that expands a monolingual TE corpus that seems promising for the task of CLTE. We built a CLTE corpus and we report a procedure that can be used to create a CLTE corpus in any pair of languages. We also report the results obtained in our experiments with the three-way classification task for CLTE and we show that this result outperform the average score of RTE (Recognizing Textual Entailment) systems. Finally, we find that using WordNet as the only source of lexical-semantic knowledge it is possibly to build a system for CLTE, which achieves comparable results with the average score of RTE systems for both two-way and three-way tasks.**

Index Terms—**Cross-lingual textual entailment, textual entailment, WordNet, bilingual textual entailment corpus.**

I. Introduction

THE objective of the Recognizing Textual Entailment (RTE) task [1] is determining whether the meaning of a hypothesis *H* can be inferred from a text *T*. Thus, we say that *T entails H*, if a person reading *T* would infer that *H* is most likely true.

Therefore, this definition assumes common human understanding of language and common background knowledge. Below, we provide an example of a *T-H* pair:

T= "*Dawson is currently a Professorial Fellow at the University of Melbourne, and an Adjunct Professor at Monash University*".

H= "*Dawson teaches at Monash University*".

In that context, Cross-Lingual Textual Entailment has been recently proposed in [2] as a generalization of Textual Entailment task (also Monolingual Textual Entailment) that consists in determining if the meaning of *H* can be inferred from the meaning of *T* when *T* and *H* are in different languages.

This new task has to face more additional issues than monolingual TE. Among them, we emphasize the ambiguity, polysemy, and coverage of the resources. Another additional problem is the necessity for semantic inference across languages, and the limited availability of multilingual knowledge resources. In RTE the most common resources used are WordNet, VerbOcean, Wikipedia, FrameNet, and DIRT. From them, only WordNet and Wikipedia are available in other languages different than English, but again, naturally with problems of coverage.

However, it is interesting to remark that, from the ablation test reported on TAC2010[1][3], some RTE systems had a positive impact using such resources, but other had a negative impact, thus the important thing is the way in which the systems utilize the available knowledge resources.

In this paper, we conduct experiments for CLTE, taking English as source language and Spanish as target language in the task of deciding the entailment among multiple languages. We chose this pair of languages due to the well-known accuracy of the translation models between Spanish and English and also due to our availability of translators whose first language is Spanish. In our work, the CLTE problem is addressed by using a machine learning approach, in which all features are WordNet-based, with the aim of measuring the benefit of WordNet as a knowledge resource for the CLTE task.

We know that the coverage of WordNet is not very good for narrow domains [4], and that also provides limited coverage of proper names. However, we are interested in evaluating the effectiveness of WordNet for CLTE, because this is the most widely used in TE. Despite these limitations, our system achieves a performance above the average score, and provides a promising direction for this line of research.

Thus, we tested a MLP and SVM classifier over two and three way decision tasks. Our focus to CLTE is based on free online (web) machine translation systems, so we chose Microsoft Bing Translator[2], because it has a good efficiency when translating English to Spanish or vice-versa, and also because provides a wide range of language pairs for translation. In addition, we use Google Translate[3], because his high efficiency has been tested in other NLP tasks [5] and [6].

This decoupled approach between Textual Entailment and Machine Translation has several advantages, such as taking

Manuscript received July 1, 2011. Manuscript accepted for publication October 2, 2011.

J. Castillo is with National University of Cordoba - FaMAF, Cordoba, Argentina and also with the National Technological University-Regional Faculty of Cordoba, Argentina (email: jotacastillo@gmail.com).

M. Cardenas is with the National Technological University-Regional Faculty of Cordoba, Argentina (email: ing.marinacardenas@gmail.com).

[1] http://www.nist.gov/tac/2010/RTE/index.html
[2] http://www.microsofttranslator.com/
[3] http://translate.google.com/

benefits of the most recent advances in machine translation, the ability to test the efficiency of different MT systems, as well as the ability to scale the system easily to any language pair.

Our approach is similar to that described in [2], because it uses a machine translation approach to CLTE. But, while they use an English-French CLTE engine with the TE engine based on edit distance algorithms, in contrast, our approach is English-Spanish CLTE, and it is completely based on semantics, because our TE engine only uses WordNet-based semantic similarity measures.

We also present the first results on assessing CLTE for the three-way decision task proposed by [7] for monolingual TE, with the idea of building a CLTE system whose outputs provide more precise informational distinctions of the judgments, making a three-way decision among *YES*, *NO*, and *UKNOWN*.

Additionally, to our knowledge, we present the first available bilingual entailment corpus aimed for the task of CLTE, which is released to the community.

This paper continues on Section 2 showing the creation of the CLTE datasets. Section 3 describes the system architecture. In section 4 we provide an experimental evaluation and discussion of the results achieved for CLTE in the two and three way tasks. Finally, Section 5 summarizes some conclusions and future work.

II. Creating the Dataset for CLTE

In order to perform experiments in CLTE, we first needed to create a corpus. Thus, we started creating a bilingual English-Spanish textual entailment corpus which was based on the original monolingual corpus from previous RTE Campaigns. We built a training set and a test set, both based on the technique of human-aided machine translation.

A. Training Set

In our experiments, we built three training sets that were generated according to the following procedure.

First, we started by selecting the original RTE3 development set, and then the hypothesis was translated from English into Spanish, using Microsoft Bing Translator as machine translation system. As a result, we generated the dataset denoted by RTE3_DS_ENtoSP.

Second, all hypotheses H are manually classified in one of three classes: Good, Regular and Bad, according to the following heuristic definition:

Good: One hypothesis *H* is classified as *Good* if its meaning is perfectly understandable for a native Spanish speaker and has the same meaning as the original hypothesis *H* that belongs to the RTE3 dataset.

Regular: One of the hypotheses *H* is classified as *Regular* if its meaning is understandable for a native Spanish speaker with little effort, or if it contains less than three syntactic errors, and it has the same meaning as the original hypothesis *H* that belongs to the RTE3 dataset.

Bad: One hypothesis *H* is classified as *Bad* if its meaning is not comprehensible to a native Spanish speaker, or has three or more syntactic errors, or if its meaning is different from the original hypothesis *H* that belongs to the RTE3 dataset.

The above procedure involved the participation of three translators whose native language is Spanish, and the classification decision was obtained from a consensus of the translators themselves. For convenience, we say that a *T-H* pair belongs to any of the above categories if the hypothesis *H* belongs to one of them. As a result, we obtained a sets of *T-H* pairs, which are denoted as RTE3_DS_ENtoSP to indicate that the dataset is composed by *T-H_Sp* pairs, where the hypothesis *H_Sp* is the translated version to Spanish from the original hypothesis *H*, and here we adopted the notation $\text{RTE3\_DS\_ENtoSP} = \{Bad\} \cup \{Regular\} \cup \{Good\}$. In a similar way, for those *T-H* pairs classified as *Good* or *Regular*, we generated the dataset: $\text{RTE3\_DS\_ENtoSP\_Good} + \text{RegPairs} = \text{RTE3\_DS\_ENtoSP} - \{Bad\}$, and finally, for those *T-H* pairs classified as *Good*, we generated the dataset: $\text{RTE3\_DS\_ENtoSP\_Good} = \text{RTE3\_DS\_ENtoSP} - \{Bad\} - \{Regular\}$.

TABLE I
EXAMPLES OF THE CLASSIFIED PAIRS

Pair ID	CLASS	Hypothesis	Comment
454	BAD	En 1945, se *eliminó* una bomba atómica sobre Hiroshima.	Wrong verb.
537	BAD	El *faro de faros* estaba situado en Alejandría.	Wrong NER.
788	BAD	Los miembros de Democrat tenían expedientes de votación fuertes de la pequeña empresa.	Don't make sense.
766	REG	Molly Walsh planea *parar el comprar* de los productos genéricos.	*parar de comprar*
18	Good	La aspirina previene la hemorragia gastrointestinal.	
756	Good	Las píldoras contaminadas contuvieron fragmentos del metal.	

The use of these training sets is motivated by the need of assessing the impact of automatic translations and manual translations performed by native Spanish speakers in the task of CLTE. Also, we are especially interested in measuring the effect of the pairs classified as *Bad* in the overall accuracy of the system.

As result, the RTE3_TS_ENtoSP_Good dataset is composed by 542 pairs, the $\text{RTE3\_TS\_ENtoSP} + \text{RegPair}$ dataset is composed by 704 pairs, and the RTE3_TS_ENtoSP dataset is composed by 800 pairs.

Table 1 illustrates some examples of the pairs classified as *Good*, *Bad* and *Regular*. When the hypothesis belongs to the class *Bad*, it is provided the justification of the human translators.

B. Test Set

In test set, we conducted a separate classification process for each annotator. The reason for this is that we are interested in assessing the agreement between the annotators on the test set built. Thus, each hypothesis *H* of the dataset was judged as *Good*, *Regular* or *Bad*, following the previous definition. We

note that pairs on which the annotators disagreed were filtered-out of the class *Good*.

We started selecting the original RTE3 test set, and then the hypothesis is translated from English into Spanish. Thus, the test set named RTE3_TS_ENtoSP is created.

First, three annotators judged each pair of the RTE3_TS_ENtoSP testset generated by Google Translate. Then, we applied the Fleiss' kappa statistical measure with the aim of assessing the reliability of agreement among the annotators. As a result, the annotators agreed in 82% of their judgment, and disagreed in 18% which corresponded to Kappa level of 0.68, regarded as substantial agreement according to [8]. The disagreement was generally found when classifying a hypothesis *H* as Regular, due to the fact that some errors in *H* could by easily corrected and thus include *H* into the class *Good*. Whereas other times, the hypothesis *H* presented some errors that justified the inclusion to the class *Bad*, for one annotator, but it was classified as *Regular* according to the criteria of another annotator. We also remark that the classes *Good* and *Bad* has high degree of agreement among annotators.

For that reason, we filtered-out the pairs classified as *Regular*, eliminating about 19% from the original pairs, and then we removed the pairs classified as *Bad*, which is an additional elimination of 10% and it is because we suppose that these pairs are not useful for inference purposes. As result, we built the dataset RTE3_TS_ENtoSP_Good. Furthermore, one annotator performed a final proofreading editing the dataset. Finally, this corpus is composed by 558 pairs, which represent a 69% of the original dataset.

In the experiments, we adopted the notation: $\text{RTE3\_TS\_ENtoSP} = \{Bad\} \cup \{\text{Regular}\} \cup \{Good\}$, and $\text{RTE3\_TS\_ENtoSP\_Good} = \text{RTE3\_TS\_ENtoSP} - \{Bad\} - \{Regular\}$.

III. System Architecture

Our system is based on a machine learning approach for CLTE. The system produces feature vectors for all datasets defined in the previous section. We experimented with SVM and MLP classifiers because of their well known performance in natural language applications. The architecture of the system is shown in Figure 1.

From Figure 1 we can see that two Online Machine Translation systems are used. Also, we note that an adaptation layer has been built in order to convert a bilingual TE task into a monolingual TE task. The datasets created on Section 2.2 are required to be in bilingual English-Spanish as inputs to the CLTE layer. In opposite, the other datasets are in monolingual English-English.

This is because some of them are used at the level of CLTE layer, and other are used at the TE level.

In all experiments it was necessary a bilingual test set in English-Spanish language.

We used the following training sets: RTE3-4C[4], and RTE4-4C[4], as proposed by the authors in [9] in order to extend the RTE data sets by using machine translation engines following a variation of the round trip translation technique. We remark that all corpus used in this paper are available to the community[4].

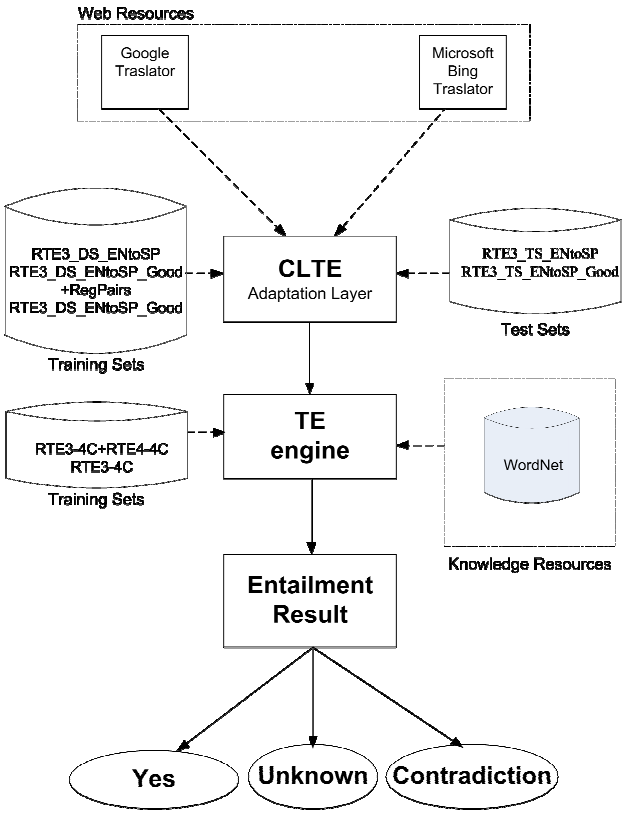

Fig. 1. General architecture of the system.

Round trip translation is defined as the process of starting with an *S* (string in English) and translating it into a foreign language *F(S)* (for example Spanish) and finally back into the English source language $F^{-1}(S)$. The Spanish language was chosen as the intermediate language, and Microsoft Bing Translator as the only MT system in this process. It was built upon the idea of providing a tool to increase the corpus size aiming to acquire more semantic variability.

The expanded corpus is denoted RTE3-4C and the three-way task is composed of: 340 pairs *Contradiction*, 1520 pairs *Yes*, and 1114 pairs *Unknown*. Thus, the two-way task is composed of: 1454 pairs *No*(No Entailment), and 1520 pairs *Yes*(Entailment). On the other hand, the RTE4-4C dataset has the following composition: 546 pairs *Contradiction*, 1812 pairs *Entailment*, and 1272 pairs *Unknown*. Therefore, in the two-way task, there are 1818 pairs *No* and 1812 pairs *Yes* in this dataset.

The sign "+" represents the union operation of sets and "4C" means "four combinations" denoting that the dataset was generated using the algorithm to expand datasets [9] and using only one Translator engine.

In addition, we also converted the three-way corpus into only two classes: *Yes* (Entailment), and *No* (No Entailment). For this purpose, both *Contradiction* and *Unknown* classes

[4]http://www.investigacion.frc.utn.edu.ar/mslabs/~jcastillo/Sagan-test-suite/

were conflated and retagged as the class *No* Entailment.

It is important to note that the dataset RTE3-4C+RTE4-4C+RTE3_DS_ENtoSP is not present in the Figure 1 because is a result of the union of dataset of both CLTE and TE layers.

Finally, our Textual Entailment engine utilizes eight WordNet-based similarity measures, such as proposed by the authors in [10], with the purpose of obtaining the maximum similarity between two concepts. These text-to-text similarity measures are based on the followings word-to-word similarity metrics: Resnik [11], Lin [12], Jiang & Conrath [13], Pirrò & Seco [14], Wu & Palmer [15], Path Metric, Leacock & Chodorow [16], and a semantic similarity to sentence level named SemSim [10].

A. Features

In this section we provide a brief resume of the text-text similarity measures which are the features of our system.

WordNet is used to calculate the semantic similarity between a T (Text) and an H (Hypothesis). The following procedure is applied:

Step 1. Perform WSD based on WordNet glosses.

Step 2. A semantic similarity matrix between *T* and *H* is defined.

Step 3. A function *Fsim* is applied to *T* and *H*.

Where the Function *Fsim* could be one of the followings seven functions over concepts *s*, and *t*:

Function 1. The Resnik similarity metric is calculated as: $RES(s,t) = IC(LCS(s,t))$, where *IC* (information content) is defined as: $IC(w) = -\log P(w)$

The function *P(w)* is the probability of selecting *w* in a large corpus, and the function *LCS(s,t)* is the least common subsume of s and t.

Function 2. The Lin similarity metric is calculated as follows:

$$LIN(s,t) = \frac{2 * IC(LCS(s,t))}{IC(s,t)}$$

Function 3. The Jiang & Conrath metric is computed as follows:

$$JICO(s,t) = \frac{1}{IC(s) + IC(t) - 2 * IC(LCS(s,t))}$$

Function 4. The Pirro & Seco (PISE) similarity metric is computed as follows:

$$PISE(s,t) = \begin{cases} 3 * IC(msca(s,t)) - IC(s) - IC(t), if\ s \neq t \\ 1 \qquad\qquad\qquad\qquad\quad , if\ s = t \end{cases}$$

The function *msca* is the most specific common abstraction value for the two given synsets (Lucene documents).

Function 5. The Wu & Palmer measure is computed as follows:

$$WUPA(C_1(s), C_2(t)) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3}$$

Where: C_1 and C_2 are the synsets to which *s* and *t* belong, respectively. C_3 is the least common superconcept of C_1 and C_2. N_1 is the number of nodes of the path from C_1 to C_3. N_2 is the number of nodes of the path from C_2 to C_3. N_3 is the number of nodes on the path from C_3 to root.

Function 6. The metric Path is reciprocal to the length of the shortest path between 2 synsets. Note that we count the 'nodes' (synsets) in the path, not the links. The allowed POS types are nouns and verbs.

$$PA(s,t) = Min_i(PathLength_i(s,t))$$

where: $PathLength_i(s,t)$ gives the length of the i-Path between *s* and *t*.

Function 7. The Leacock & Chodorow metric finds the path length between s and t in the "is-a" hierarchy of WordNet, and is computed as follows:

$$LECH(C_1(s), C_2(t)) = -\log\left(\frac{Min_i(PathLength_i(s,t)}{2 * D}\right)$$

where: *D* = is the maximum depth of the taxonomy (considering only nouns and verbs).

Step 4. Finally, the string similarity between two lists of words is reduced to the problem of bipartite graph matching by using the Hungarian algorithm over this bipartite graph. Then, we find the assignment that maximizes the sum of ratings of each token. Note that each graph node is a token/word of the list.
At the end, the final score is calculated by:

$$finalscore = \frac{\sum_{s \in T, t \in H} opt(Fsim(s,t))}{Max\ (Length(T), Length\ (H))}$$

where: *opt(F)* is the optimal assignment in the graph.
Length (T) is the number of tokens in *T*, *Length (H)* is the number of tokens in *H*, and

$$F_{sim} \in \{RES, LIN, JICO, PISE, WUPA, PA, LECH\}$$

Finally, note that the partial influence of each of the individual similarities will be reflected on the overall similarity.

Function 8. Additionally, the SemSim metric is defined and calculated as follows:
Step 1. Perform WSD based on WordNet definitions.
Step 2. Compute a semantic similarity matrix between words in *T* and *H*, using only synonym and hyperonym relationship. The Breadth First Search algorithm is used over these tokens. Then, the semantic similarity between two words/concepts *s* and *t*, is computed as:

$$Sim\ (s,t) = 2 \times \frac{Depth\ (LCS(s,t))}{Depth\ (s) + Depth\ (t)}$$

where: *Depth(s)* is the shortest distance from the root node to the current node.
Step 3. In this step the Function 8 is computed. Thus, in order to obtain the final score, the matching average between two sentences *T* and *H* is calculated as follows:

$$SemSim(T,H) = 2 \times \frac{Match\ (T,H)}{Length\ (T) + Length\ (H)}$$

Finally, this procedure produces eight WordNet-based semantic similarity measures, which have been tested over monolingual textual entailment [10] achieving results that outperformed the average accuracy of the RTE systems.

IV. Results and Discussion

In this section, we test the system to predict the following test sets: RTE3_TS_ENtoSP and RTE3_TS_ENtoSP_Good. In the experiments performed we used the training sets given below:

- RTE3_DS_ENtoSP,
- RTE3_DS_ENtoSP_Good+RegPairs, and
- RTE3_DS_ENtoSP_Good.

Additionally, we utilize the RTE3-4C, and RTE3-4C+RTE4-4C datasets.

We generated a feature vector for every *T-H* pair with both training and test sets. The feature vector is composed of the following eight components: F_{RES}, F_{LIN}, F_{JICO}, F_{PISE}, F_{WUPA}, F_{PA}, F_{LECH}, and SemSim. The achieved results are shown in Table 2 and Table 3.

Results reported in both tables show that we achieved the best performance, or nearly the best, with the dataset RTE3-4C+RTE4-4C in the majority of the cases.

It is interesting to note that our best result in the two-way task is obtained to predict the RTE3_TS_ENtoSP_Good test set, which is actually the realistic case, because this dataset contains only pairs validated by humans. On the other hand, the test set RTE3_TS_ENtoSP contains *BAD* pairs, and we obtained results comparables to those obtained with the previous case.

On the contrary, in the case of three-way task, the highest results are achieved considering RTE3_TS_ENtoSP as test set.

In both cases, the difference found when predicting RTE3_TS_ENtoSP and RTE3_TS_ENtoSP_Good is not statistical significant.

Surprisingly, the worse results in all the cases were obtained with the RTE3_DS_ENtoSP_Good as training set. This can be caused by the size of this dataset, which is composed by only 542 pairs.

As we previously note, the datasets RTE3-4C and RTE4-4C have been created for monolingual textual entailment, however the system is able to use these datasets because of our decoupled approach for CLTE. Thus, this result suggests that the corpus used on monolingual task improves the result of the CLTE system.

As a term of comparison, in the RTE3 Challenge [17] the average score achieved in the two-way task for the monolingual textual entailment was 62.37% of accuracy reached by the competing systems, which is 0.75% and 1.13% below our accuracy levels of 63.12% and 63.5% obtained with the SVM classifier and using the RTE3-4C+RTE4-4C and RTE3-4C+RTE4-4C+ RTE3_DS_ENtoSP as the training sets, but not resulting in a significant statistical difference.

In the RTE4 Challenge, the average score achieved in the three-way task was 50.65%, and thus our system outperforms on 9.63% when using SVM and RTE3-4C+RTE4-4C as training set, which is a significant statistical difference, although these sets are not actually comparable.

Although the elements belonging to the class *Bad* are included in RTE3_DS_ENtoSP, surprisingly, better performances are achieved in comparison with other data sets with neither *Regular* nor *Bad* pairs. The *T-H* pairs included in the set *Bad* have some syntax errors and, even more, are not understandable by the translators. However, many of the words "w" in the text *T* are also present in the hypothesis the *H* as "w", or are present as synonyms of "w", which increases the semantic correspondence between the *T-H* pair. This could be a reason for the increasing in efficiency when using RTE3_DS_ENtoSP as training set.

Interestingly, if we analyze only the size of data sets, we see that the larger the training set, the greater the efficiency gains. This highlights the need for larger datasets for the purpose of building more accurate models. It is also showed by the best accuracy that is found in our system when using the expanded dataset RTE3-4C+RTE4-4C.

TABLE II
ACCURACY OBTAINED CONSIDERING RTE3_TS_ENTOSP AND RTE3_TS_ENTOSP_GOOD AS TEST SET IN THE TWO-WAY TASK

Datasets	RTE3_TS_ENtoSP		RTE3_TS_ENtoSP_ Good	
Classifiers	2-way MLP Classifier	2-way SVM Classifier	2-way MLP Classifier	2-way SVM Classifier
RTE3_DS_ENtoSP	59.75	62.12	60.46	61.53
RTE3_DS_ENtoSP_Good +RegPairs	58.37	60.62	59.39	61.53
RTE3_DS_ENtoSP_Good	57.62	58.12	57.96	61.53
RTE3-4C+RTE4-4C	60.37	**63.12**	**63.32**	62.96
RTE3-4C	**62.62**	61.75	62.61	62.43
RTE3-4C+RTE4-4C+ RTE3_DS_ENtoSP	62.62	62.25	62.79	**63.50**

TABLE III
ACCURACY OBTAINED CONSIDERING RTE3_TS_ENTOSP AND RTE3_TS_ENTOSP_GOOD AS TEST SET IN THE THREE-WAY TASK

Datasets	RTE3_TS_ENTOSP		RTE3_TS_ENtoSP_ Good	
Classifiers	3-way MLP Classifier	3-way SVM Classifier	3-way MLP Classifier	3-way SVM Classifier
RTE3_DS_ENtoSP	57.96	58.31	57.96	**58.31**
RTE3_DS_ENtoSP_Good +RegPairs	58.49	**58.85**	58.14	56.35
RTE3_DS_ENtoSP_Good	54.75	56.62	55.09	55.28
RTE3-4C+RTE4-4C	**60.28**	58.14	58.32	58.14
RTE3-4C	58.75	57.50	58.32	58.14
RTE3-4C+RTE4-4C+ RTE3_DS_ENtoSP	59.87	57.50	**59.57**	58.14

V. Conclusion

From our experiments, we conclude that a promising algorithm to expand an RTE Corpus yielded significant statistical differences when predicting RTE test sets. We also show that although WordNet is not enough to build a competitive TE system, an average score could be reached or outperformed for the CLTE task.

As a further contribution, our experiments suggest that using the expanded method for the corpus can increase the accuracy of CLTE systems, in both two-way and three-way tasks. All results obtained in these tasks are comparable (or outperformed) with the average score of existing RTE systems. As additional result, we present the first CLTE corpus, and a procedure to create a corpus with the technique of human-aided machine translation, which also could be used to create a bilingual TE corpus in any language pairs. This corpus reaches an inter-annotator agreement corresponding to Kappa level of 0.68, regarded as substantial agreement.

Furthermore, the results obtained for the three-way task in CLTE outperforms the score of an average system by 9.63% accuracy when predicting the RTE3_TS_ENtoSP dataset.

Our future work will address the incorporation of additional knowledge resources and will incorporate additional lexical similarities features and semantic resources and assess the improvements they may yield. Finally, we aim at releasing additional CLTE corpus to the community in the future.

References

[1] L. Bentivogli, I. Dagan, H. Dang, D. Giampiccolo, and B. Magnini, "The Fifth PASCAL RTE Challenge," in *Proceedings of the Text Analysis Conference*, 2009.

[2] Y. Mehdad, M. Negri, and M. Federico, "Towards Cross-Lingual Textual entailment," in *Proceedings of the 11th NAACL HLT*, 2010.

[3] L. Bentivogli, P. Clark, I. Dagan, H. Dang, D. Giampiccolo, "The Sixth Pascal Recognizing Textual Entailment Challenge," in *Proceedings of Textual Analysis Conference*, NIST, Maryland USA, 2010.

[4] R. Richardson and A. Smeaton, "Using WordNet in a Knowledge-Based Approach to Information Retrieval," *Techn. Report Working Paper*: CA-0395, Dublin City University, Dublin, Ireland, 1995.

[5] J. Marlow, P. Clough, J. Recuero, and J. Artiles, "Exploring the Effects of Language Skills on Multilingual Web Search," in *Proceedings of the 30th European Conference on IR Research (ECIR'08)*, Glasgow, UK. LNCS, Volume 4956, Springer, Heidelberg, 2008, pp. 126−137.

[6] J. Lilleng and S. Tomassen, "Cross-lingual information retrieval by feature vectors", *NLDB 2007, LNCS*, pp. 229−239, 2007.

[7] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan, "The Third PASCAL Recognizing Textual Entailment Challenge," in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic, 2007.

[8] J. Landis and G. Koch, "The measurements of observer agreement for categorical data," *Biometrics*, 33:159–174, 1997.

[9] J. Castillo, "Using Machine Translation to expand a Corpus in Textual Entailment," in *Proceedings of the 7th ICETAL*, Reykjavik, Iceland. LNCS, vol. 6233, Springer, Heidelberg, 2010, pp. 97–102.

[10] J. Castillo, "A Semantic Oriented Approach to Textual Entailment using WordNet-based Measures," in *Proceedings of the MICAI 2010*, Pachuca, Mexico, LNCS, vol. 6437, Springer, Heidelberg, 2010, pp. 44–55.

[11] P. Resnik, "Information Content to Evaluate Semantic Similarity in a Taxonomy," in *Proceedings of IJCAI 1995*, 1995, pp. 448−453.

[12] D. Lin, "An Information-Theoretic Definition of Similarity," in *Proceedings of Conference on Machine Learning*, 1997, pp. 296–304.

[13] J. Jiang and D. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," in *Proceedings of the ROCLING X*, 1997.

[14] G. Pirrò and N. Seco, "Design, Implementation and Evaluation of a New Similarity Metric Combining Feature and Intrinsic Information Content," *ODBASE 2008*, Springer LNCS, 2008.

[15] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in *Proceedings of the 32nd ACL*, 1994.

[16] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," in *WordNet: An Electronic Lexical Database*, MIT Press, pp. 265−283, 1998.

[17] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan, "The Third PASCAL Recognizing Textual Entailment Challenge," in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic, 2007.

[18] T. Pedersen, S. Patwardhan, and J. Michelizzi, "WordNet::Similarity - Measuring the Relatedness of Concepts," in *Proceedings of the AAAI-04*, 2004.

[19] C. Quirk, C. Brockett, and W. Dolan, "Monolingual Machine Translation for Paraphrase Generation," in *Proceedings of the ACL-HLT*, 2004.