BUAP: N-gram based Feature Evaluation for the Cross-Lingual Textual Entailment Task

Darnes Vilariño, David Pinto, Saúl León, Yuridiana Alemán, Helena Gómez-Adorno
Benemérita Universidad Autónoma de Puebla
Faculty of Computer Science
14 Sur y Av. San Claudio, CU
Puebla, Puebla, México
{darnes, dpinto, saul.leon, candy.aleman, helena.gomez}@cs.buap.mx

Abstract

This paper describes the evaluation of different kinds of textual features for the Cross-Lingual Textual Entailment Task of SemEval 2013. We have counted the number of N-grams for three types of textual entities (character, word and PoS tags) that exist in the pair of sentences from which we are interested in determining the judgment of textual entailment. Difference, intersection and distance (Euclidian, Manhattan and Jaccard) of N-grams were considered for constructing a feature vector which is further introduced in a support vector machine classifier which allows to construct a classification model. Five different runs were submitted, one of them considering voting system of the previous four approaches. The results obtained show a performance below the median of six teams that have participated in the competition.

1 Introduction

The cross-lingual textual entailment (CLTE), recently proposed by (Mehdad et al., 2012) and (Mehdad et al., 2011), is an extension of the textual entailment task (Dagan and Glickman, 2004). Formally speaking, given a pair of topically related text fragments ($T1$ and $T2$ which are assumed to be TRUE statements) written in different languages, the CLTE task consists of automatically annotating it with one of the following entailment judgments:

- bidirectional ($T1 \rightarrow T2$ & $T1 \leftarrow T2$): the two fragments entail each other (semantic equivalence);
- forward ($T1 \rightarrow T2$ & $T1 \nleftarrow T2$): unidirectional entailment from $T1$ to $T2$;
- backward ($T1 \nrightarrow T2$ & $T1 \leftarrow T2$): unidirectional entailment from $T2$ to $T1$;
- no entailment (T1 $\nrightarrow T2$ & $T1 \nleftarrow T2$): there is no entailment between $T1$ and $T2$ in both directions;

The Cross-lingual datasets evaluated were available for the following language combinations ($T1$-$T2$):

- Spanish-English (SPA-ENG)
- German-English (DEU-ENG)
- Italian-English (ITA-ENG)
- French-English (FRA-ENG)

In this paper we describe the evaluation of different features extracted from each pair of topically related sentences. N-grams of characters, words and PoS tags were counted with the aim of constructing a representative vector for each judgment entailment (FORWARD, BACKWARD, BI-DIRECTIONAL or NO-ENTAILMENT). The resulting vectors were fed into a supervised classifier based on Support Vector Machines (SVM)[1] which attempted to construct a classification model. The description of the features and the vectorial representation is given in Section 2. The obtained results are shown and discussed in Section 3. Finally, the findings of this work are given in Section 4.

[1] We have employed the implementation of the Weka tool (Hall et al., 2009).

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 124–127, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

2 Experimental Setup

We have considered the task as a classification problem using the pivot approach. Thus, we have translated[2] each pair to their corresponding language in order to have two pairs of sentences written in the same language. Let $Pair(T1, T2)$ be the original pair of topically related sentences. Then, we have obtained the English translation of $T1$, denoted by $T3$, which will be aligned with $T2$. On the other hand, we have translated $T2$ to the other language (Spanish, German, Italian or French), denoted by $T4$, which will be aligned with $T1$. The two pairs of sentences, $Pair(T2, T3)$ (English) and $Pair(T1, T4)$ (other language), are now written in the same language, and we can proceed to calculate the textual features we are interested in.

The features used to represent both sentences are described below:

- N-grams of characters, with $N = 2, \cdots, 5$.
- N-grams of words, with $N = 2, \cdots, 4$.
- N-grams of PoS tags, with $N = 2, \cdots, 4$.
- Euclidean measure between each pair of sentences ($Pair(T1, T4)$ and $Pair(T2, T3)$).
- Manhattan measure between each pair of sentences ($Pair(T1, T4)$ and $Pair(T2, T3)$).
- Jaccard coefficient, expanding English terms in both sentences, $T2$ and $T3$, with their corresponding synonyms (none disambiguation process was considered).

The manner we have used the above mentioned features is described in detail in the following subsections.

2.1 Approach 1: Difference operator

For each pair of sentences written in the same language, this approach counts the number of N-grams that occur in the first sentence (for instance $T1$), and do not occur in the second sentence (for instance $T4$) and viceversa. Formally speaking, the values obtained are $\overrightarrow{Pair}(T1, T2) = \{D_1, D_2, \cdots, D_k\}$, with $D_1 = |T1 - T4|$, $D_2 = |T4 - T1|$, $D_3 = |T2 - T3|$, $D_4 = |T3 - T2|$, $\cdots$. This vector is calculated for all the possible values of N for each type of N-gram, i.e., character, word and PoS tag. The cardinality of $\overrightarrow{Pair}(T1, T2)$ will be 34, that is, 16 values when the N-grams of characters are considered, 12 values with word N-grams, and 6 values when the PoS tag N-grams are used.

Table 1: Classes considered in the composition of binary classifiers

Class 1	Class 2
BACKWARD	OTHER
BI-DIRECTIONAL	OTHER
FORWARD	OTHER
NO-ENTAILMENT	OTHER
BACKWARD & BI-DIRECTIONAL	OTHER
BACKWARD & FORWARD	OTHER
BACKWARD & NO-ENTAILMENT	OTHER
BI-DIRECTIONAL & NO-ENTAILMENT	OTHER
FORWARD & BI-DIRECTIONAL	OTHER
FORWARD & NO-ENTAILMENT	OTHER

The vectors obtained are labeled with the corresponding tag in order to construct a training dataset which will be further used to feed a multiclass classifier which constructs the final classification model. In this case, the system will directly return one of the four valid entailment judgments (i.e. forward, backward, bidirectional, no_entailment).

2.2 Approach 2: Difference and Intersection operators

This approach enriches the previous one, by adding the intersection between the two sentences of each pair. In a sense, we have considered all the features appearing in the pair of sentences. In this case, the total number of features extracted, i.e., the cardinality of the $\overrightarrow{Pair}(T1, T2)$ vector is 51.

2.3 Approach 3: Metaclassifier

In this approach, we have constructed a system which is a composition of different binary classification models. The binary judgments were constructed considering the classes shown in Table 1.

The approach 2 was also considered in this composition generating a total of 11 models. 10 of them are based on the features used by Approach 1, and the last one is based on the features used by Approach 2. The result obtained is a vector which tells whether or not a pair is judged to have some kind of textual entailment or not (the OTHER class). This

[2] For this purpose we have used Google Translate

vector is then labeled with the correct class obtained from the gold standard (training corpus) for automatically obtaining a decision tree which allows us to determine the correct class. Thus, the different outputs of multiple classifiers are then introduced to another supervised classifier which constructs the final classification model.

2.4 Approach 4: Distances measures

This approach is constructed by adding five distance values to the Approach 2. These values are calculated as follows :

- The Euclidean distance between $T2$ and $T3$, and between $T1$ and $T4$. We have used the frequency of each word for constructing a representative vector of each sentence.
- The Manhattan distance between $T2$ and $T3$, and between $T1$ and $T4$. We have used the frequency of each word for constructing a representative vector of each sentence.
- A variant of the Jaccard's Coefficient that consider synonyms (Carrillo et al., 2012). Since we have only obtained synonyms for the English language, this measure was only calculated between $T2$ and $T3$.

Therefore, the total number of features of the $\overrightarrow{Pair}(T1, T2)$ vector is 56.

2.5 Approach 5: Voting system

With the results of the previous four models, we prepared a voting system which uses the majority criterion (3 of 4).

3 Experimental results

The results obtained in the competition are presented and discussed in this section. First, we describe the training and test corpus, and thereafter, the results obtained with the different approaches submitted.

3.1 Dataset

In order to train the different approaches already discussed, we have constructed a training corpus made up of two datasets: the training data provided by the task organizers the task 8 of SemEval 2013 (Negri et al., 2013), and the test dataset together with the gold standard of CLTE task of SemEval 2012 (Negri et al., 2011). Thus, the training corpus contains 4000 sentence pairs. The test set provided in the competition contains 2000 sentence pairs. The corpus is balanced, with 1000 pairs for each language in the training dataset, whereas, 500 pairs are given in the test set for each language (see Table 2).

Table 2: Description of the dataset

Languages	**Training**	**Test**
SPA-ENG	1000	500
DEU-ENG	1000	500
ITA-ENG	1000	500
FRA-ENG	1000	500
Total	**4000**	**2000**

3.2 Results

In Table 3 we can see the results obtained by each one of the five approaches we submitted to the competition. Each approach has been labeled with the prefix "BUAP-R" for indicating the approach used by each submitted run. For instance, the BUAP-R1 run corresponds to the approach 1 described in the previous section. As can be seen, the behavior of the five approaches is quite similar, which we consider it is expected because the underlying methodology employed is almost the same for all the approaches. With exception of the pair of sentences written in SPA-ENG in which the best approach was obtained by the BUAP-R5 run, the approach 4 outperformed the other appproaches. We believe that this has been a result of introducing measures of similarity between the two sentences and their translations. In this table it is also reported the Highest, Average, Median and Lowest values of the competition. The results we obtained are under the Median but outperformed the results of two teams in the competition.

With the purpose of analyzing the behavior of the approach 4 in each one of the entailment judgments, we have provided the results obtained in Table 4. There we can see that the BACKWARD class is the easiest one for being predicted, independently of the language. The second easiest class is FORWARD, followed by NO-ENTAILMENT. Also we can see that the BI-DIRECTIONAL class is the one that produce more confusion, thus leading to obtain a lower performance than the other ones.

Table 3: Overall statistics obtained in the Task-8 of SemEval 2013

RUN	SPA-ENG	ITA-ENG	FRA-ENG	DEU-ENG
Highest	0.434	0.454	0.458	0.452
Average	0.393	0.393	0.401	0.375
Median	0.392	0.402	0.416	0.369
Lowest	0.340	0.324	0.334	0.316
BUAP-R1	0.364	0.358	0.368	0.322
BUAP-R2	0.374	0.358	0.364	0.318
BUAP-R3	0.380	0.358	0.362	0.316
BUAP-R4	0.364	**0.388**	**0.392**	**0.350**
BUAP-R5	**0.386**	0.360	0.372	0.318

Table 4: Statistics of the approach 4, detailed by entailment judgment

ENTAILMENT JUDGEMENT	SPA-ENG	ITA-ENG	FRA-ENG	DEU-ENG
BACKWARD	0.495	0.462	0.431	0.389
FORWARD	0.374	0.418	0.407	0.364
NO-ENTAILMENT	0.359	0.379	0.379	0.352
BI-DIRECTIONAL	0.277	0.327	0.352	0.317

4 Conclusions

Five different approaches for the Cross-lingual Textual Entailment for the Content Synchronization task of Semeval 2013 are reported in this paper. We used several features for determining the textual entailment judgment between two texts $T1$ and $T2$ (written in two different languages). The approach 4 proposed, which employed lexical similarity and semantic similarity in English language only was the one that performed better. As future work, we would like to include more distance metrics which allow to extract additional features of the pair of sentences topically related.

References

Maya Carrillo, Darnes Vilariño, David Pinto, Mireya Tovar, Saul León, and Esteban Castillo. Fcc: Three approaches for semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1 and 2 (SemEval 2012)*, pages 631–634, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

Ido Dagan and Oren Glickman. Probabilistic textual entailment: Generic applied modeling of language variability. In *PASCAL Workshop on Learning Methods for Text Understanding and Mining*, 2004.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. ISSN 1931-0145.

Yashar Mehdad, Matteo Negri, and Marcello Federico. Using bilingual parallel corpora for cross-lingual textual entailment. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1336–1345, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Yashar Mehdad, Matteo Negri, and Marcello Federico. Detecting semantic equivalence and information disparity in cross-lingual documents. In *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 120–124, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

M. Negri, A. Marchetti, Y. Mehdad, L. Bentivogli, and D. Giampiccolo. Semeval-2013 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, 2013.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. Divide and conquer: crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 670–679, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-11-4.