BUAP: Lexical and Semantic Similarity for Cross-lingual Textual Entailment

Darnes Vilariño, David Pinto, Mireya Tovar, Saul León, Esteban Castillo
Benemérita Universidad Autónoma de Puebla,
Faculty of Computer Science
14 Sur & Av. San Claudio, CU
Puebla, Puebla, México
{darnes, dpinto, mtovar}@cs.buap.mx
saul.ls@live.com, ecjbuap@gmail.com

Abstract

In this paper we present a report of the two different runs submitted to the task 8 of Semeval 2012 for the evaluation of Cross-lingual Textual Entailment in the framework of Content Synchronization. Both approaches are based on textual similarity, and the entailment judgment (bidirectional, forward, backward or no entailment) is given based on a set of decision rules. The first approach uses textual similarity on the translated and original versions of the texts, whereas the second approach expands the terms by means of synonyms. The evaluation of both approaches show a similar behavior which is still close to the average and median.

1 Introduction

Cross-lingual Textual Entailment (CLTE) has been recently proposed by (Mehdad et al., 2010; Mehdad et al., 2011) as an extension of the Textual Entailment task (Dagan and Glickman, 2004). Given a text (T) and an hypothesis (H) in different languages, the CLTE task consists of determining if the meaning of H can be inferred from the meaning of T. In this paper we present a report of the obtained results after submitting two different runs for the Task 8 of Semeval 2012, named "Cross-lingual Textual Entailment for Content Synchronization" (Negri et al., 2012). In this task, the Cross-Lingual Textual Entailment addresses textual entailment recognition under a new dimension (cross-linguality), and within a new challenging application scenario (content synchronization). The task 8 of Semeval 2012 may be formally defined as follows:

Given a pair of topically related text fragments (T_1 and T_2) in different languages, the task consists of automatically annotating it with one of the following entailment judgments:

- Bidirectional ($T_1 \rightarrow T_2$ & $T_1 \leftarrow T_2$): the two fragments entail each other (semantic equivalence)
- Forward ($T_1 \rightarrow T_2$ & $T1\ !\leftarrow T_2$): unidirectional entailment from T_1 to T_2
- Backward ($T_1\ !\rightarrow T_2$ & $T_1 \leftarrow T_2$): unidirectional entailment from T_2 to T_1
- No Entailment ($T_1\ !\rightarrow T_2$ & $T_1\ !\leftarrow T_2$): there is no entailment between T_1 and T_2

In this task, both T_1 and T_2 are assumed to be *TRUE* statements; hence in the dataset there are no contradictory pairs. Cross-lingual datasets are available for the following language combinations:

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

The remaining of this paper is structured as follows: Section 2 describes the two different approaches presented in the competition. The obtained results are shown and dicussed in Section 3. Finally, the findings of this work are given in Section 4.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 706–709,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

2 Experimental setup

For this experiment we have considered to tackle the CLTE task by means of textual similarity and textual length. In particular, the textual similarity is used to determine whether some kind of entailment exists or not. We have established the threshold of 0.5 for the similarity function as evidence of textual entailment. Since the two sentences to be evaluated are written in two different languages, we have translated each sentence to the other language, so that, we have two sentences in English, and two sentences in the original language (Spanish, German, Italian and French). We have used the Google translate for this purpose [1].

The corpora used in the experiments comes from a cross-lingual Textual Entailment dataset presented in (Negri et al., 2011), and provided by the task organizers. We have employed the training dataset only for adjust some parameters of the system, but the approach is knowledge-based and, therefore, it does not need a training corpus. Both, the training and test corpus contain 500 sentences for each language.

The textual length is used to determine the entailment judgment (bidirectional, forward, backward, no entailment). We have basically, assumed that the length of a text may give some evidence of the type of entailment. The decision rules used for determining the entailment judgment are described in Section 2.3.

In this competition we have submitted two different runs which differ with respect to the type of textual similarity used (lexical vs semantic). The first one, calculates the similarity using only the translated version of the original sentences, whereas the second approach uses text expansion by means of synonyms and, thereafter, it calculates the similarity between the pair of sentences.

Let T_1 be the sentence in the original language, T_2 the T_1 topically related text fragment (written in English). Let T_3 be the English translation of T_1, and T_4 the translation of T_2 to the original language (Spanish, German, Italian and French). The formal description of these two approaches are given as follows.

2.1 Approach 1: Lexical similarity

The evidence of textual entailment between $T1$ and $T2$ is calculated using two formulae of lexical similarity. Firstly, we determine the similarity between the two texts written in the source language ($SimS$). Additionally, we calculate the lexical similarity between the two sentences written in the target language ($SimT$), in this case English.

Given the limited text length of the text fragments, we have used the Jaccard coefficient as similarity measure. Eq. (1) shows the lexical similarity for the two texts written in the original language, whereas, Eq. (2) presents the Jaccard coefficient for the texts written in English.

$$simS = simJaccard(T_1, T_4) = \frac{|T_1 \cup T_4|}{|T_1 \cap T_4|} \quad (1)$$

$$simT = simJaccard(T_2, T_3) = \frac{|T_2 \cup T_3|}{|T_2 \cap T_3|} \quad (2)$$

2.2 Approach 2: Semantic similarity

In this case we calculate the semantic similarity between the two texts written in the original language ($simS$), and the semantic similarity between the two text fragments written in English ($simT$). The semantic level of similarity is given by considering the synonyms of each term for each sentence (in the original and target language). For this purpose, we have employed five dictionaries containing synonyms for the five different languages considered in the competition (English, Spanish, German, Italian, and French)[2]. In Table 1 we show the number of terms, so as the number of synonyms in average by term considered for each language.

Let $T_1 = w_{1,1}w_{1,2}...w_{1,|T_1|}$, $T_2 = w_{2,1}w_{2,2}...w_{2,|T_2|}$ be the source and target sentences, and let $T_3 = w_{3,1}w_{3,2}...w_{3,|T_3|}$, $T_4 = w_{4,1}w_{4,2}...w_{4,|T_4|}$ be translated version of the original source and target sentences, respectively. The synonyms of a given word $w_{i,k}$, expressed as $synset(w_{i,k})$, are obtained from the aforementioned dictionaries by extracting the synonyms of $w_{i,k}$. In order to obtain a better matching between the terms contained in the text fragments and the terms in the

[1] http://translate.google.com.mx/

[2] http://extensions.services.openoffice.org/en/dictionaries

Table 1: Dictionaries of synonyms used for term expansion

Language	*Terms*	*synonyms per term (average)*
English	2,764	60
Spanish	9,887	45
German	21,958	115
Italian	25,724	56
French	36,207	93

dictionary, we have stemmed all the terms using the Porter stemmer.

In order to determine the semantic similarity between two terms of sentences written in the source language ($w_{1,i}$ and $w_{4,j}$) we use Eq. (3). The semantic similariy between two terms of the English sentences are calculated as shown in Eq. (4).

$$sim(w_{1,i}, w_{4,j}) = \begin{cases} 1 & \text{if } (w_{1,i} == w_{4,j}) \mid\mid \\ & w_{1,i} \in synset(w_{4,j}) \mid\mid \\ & w_{4,j} \in synset(w_{1,i}) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$sim(w_{2,i}, w_{3,j}) = \begin{cases} 1 & \text{if } (w_{2,i} == w_{3,j}) \mid\mid \\ & w_{2,i} \in synset(w_{3,j}) \mid\mid \\ & w_{3,j} \in synset(w_{2,i}) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Both equations consider the existence of semantic similarity when the two words are identical, or when the some of the two words appear in the synonym set of the other word.

The semantic similarity of the complete text fragments T_1 and T_4 ($simS$) is calculated as shown in Eq. (5). Whereas, the semantic similarity of the complete text fragments T_2 and T_3 ($simT$) is calculated as shown in Eq. (6).

$$simS(T_1, T_4) = \frac{\sum_{i=1}^{|T_1|} \sum_{j=1}^{|T_4|} sim(w_{1,i}, w_{4,j})}{|T_1 \cup T_4|} \tag{5}$$

$$simT(T_2, T_3) = \frac{\sum_{i=1}^{|T_2|} \sum_{j=1}^{|T_3|} sim(w_{2,i}, w_{3,j})}{|T_2 \cup T_3|} \tag{6}$$

2.3 Decision rules

Both approches used the same decision rules in order to determine the entailment judgment for a given pair of text fragments (T_1 and T_2). The following algorithm shows the decision rules used.

Algorithm 1.

```
    If     $|T_2| < |T_3|$ then
           If ($simT > 0.5$ and $simS > 0.5$)
           then forward
ElseIf     $|T_2| > |T_3|$ then
           If ($simT > 0.5$ and $simS > 0.5$)
           then backward
ElseIf     ($|T_1| == |T_4|$ and $|T_2| == |T_3|$) then
           If ($simT > 0.5$ and $simS > 0.5$)
           then bidirectional
  Else     no entailment
```

As mentioned above, the rules employed the lexical or semantic textual similarity, and the textual length for determining the textual entailment.

3 Results

In Table 2 we show the overall results obtained by the two approaches submitted to the competition. We also show the highest, lowest, average and median overall results obtained in the competition.

	SPA-ENG	*ITA-ENG*	*FRA-ENG*	*DEU-ENG*
Highest	0.632	0.566	0.57	0.558
Average	0.407	0.362	0.366	0.357
Median	0.346	0.336	0.336	0.336
Lowest	0.266	0.278	0.278	0.262
BUAP_run1	0.35	0.336	0.334	0.33
BUAP_run2	0.366	0.344	0.342	0.268

Table 2: Overall statistics obtained in the Task 8 of Semeval 2012

The runs submitted perform similar, but the semantic approach obtained a slightly better performance. The two results are above the median but below the average. We consider that better results may be obtained if the two features used (textual similarity and textual length) were introduced into a supervised classifier, so that, the decision rules were approximated on the basis of a training dataset, instead of the empirical setting done in this work. Future experiments will be carried out in this direction.

4 Discussion and conclusion

Two different approaches for the Cross-lingual Textual Entailment for Content Synchronization task of Semeval 2012 are reported in this paper. We used two features for determining the textual entailment judgment between two texts T_1 and T_2 (written in two different languages). The first approach proposed used lexical similarity, meanwhile the second used semantic similarity by means of term expansion with synonyms.

Even if the performance of both approaches is above the median and slighly below the average, we consider that we may easily improve this performance by using syntactic features of the text fragments. Additionally, we are planning to integrate some supervised techniques based on decision rules which may be trained in a supervised dataset. Future experiments will be executed in this direction.

Acknowledgments

This project has been partially supported by projects CONACYT #106625, #VIAD-ING11-II and VIEP #PIAD-ING11-II.

References

Ido Dagan and Oren Glickman. 2004. Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. In *Learning Methods for Text Understanding and Mining*, January.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2010. Towards Cross-Lingual Textual Entailment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 321–324, Los Angeles, California, June. Association for Computational Linguistics.

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

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

M. Negri, A. Marchetti, Y. Mehdad, L. Bentivogli, and D. Giampiccolo. 2012. Semeval-2012 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.