UAlacant: Using Online Machine Translation for Cross-Lingual Textual Entailment

Miquel Esplà-Gomis and **Felipe Sánchez-Martínez** and **Mikel L. Forcada**
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain
{mespla,fsanchez,mlf}@dlsi.ua.es

Abstract

This paper describes a new method for cross-lingual textual entailment (CLTE) detection based on machine translation (MT). We use sub-segment translations from different MT systems available online as a source of cross-lingual knowledge. In this work we describe and evaluate different features derived from these sub-segment translations, which are used by a support vector machine classifier to detect CLTEs. We presented this system to the SemEval 2012 task 8 obtaining an accuracy up to 59.8% on the English–Spanish test set, the second best performing approach in the contest.

1 Introduction

Cross-lingual textual entailment (CLTE) detection (Mehdad et al., 2010) is an extension of the textual entailment (TE) detection (Dagan et al., 2006) problem. TE detection consists of finding out, for two text fragments T and H in the same language, whether T entails H from a semantic point of view or not. CLTE presents a similar problem, but with T and H written in different languages.

During the last years, many authors have focused on resolving TE detection, as solutions to this problem have proved to be useful in many natural language processing tasks, such as question answering (Harabagiu and Hickl, 2006) or machine translation (MT) (Mirkin et al., 2009; Padó et al., 2009). Therefore, CLTE may also be useful for related tasks in which more than one language is involved, such as cross-lingual question answering or cross-lingual information retrieval. Although CLTE detection is a relatively new problem, it has already been tackled. Mehdad et al. (2010) propose to use machine translation (MT) to translate H from L_H, the language of H, into L_T, the language of T, and then use any of the state-of-the-art TE approaches. In a later work (Mehdad et al., 2011), the authors use MT, but in a more elaborate way. They train a phrase-based statistical MT (PBSMT) system (Koehn et al., 2003) translating from L_H to L_T, and use the translation table obtained as a by-product of the training process to extract a set of features which are processed by a support vector machine classifier (Theodoridis and Koutroumbas, 2009, Sect. 3.7) to decide whether T entails H or not. Castillo (2011) discusses another machine learning approach in which the features are obtained from semantic similarity measures based on WordNet (Miller, 1995).

In this work we present a new approach to tackle the problem of CLTE detection using a machine learning approach, partly inspired by that of Mehdad et al. (2011). Our method uses MT as a source of information to detect semantic relationships between T and H. To do so, we firstly split both T and H into all the possible sub-segments with lengths between 1 and L, the maximum length, measured in words. We then translate the set of sub-segments from T into L_H, and vice versa, and collect all the sub-segment pairs in a single set. We claim that when T-side sub-segments match T and their corresponding H-side sub-segments match H, this reveals a semantic relationship between them, which can be used to determine whether T entails H or not. Note that MT is used as a *black box*, i.e. sub-segment translations may be collected from any MT system, and that our approach could even use any other sources of bilingual sub-sentential information. It is even possible to combine different MT systems as we do in our experiments. This is a key point of our work, since

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 472–476,
Montréal, Canada, June 7-8, 2012.

it uses MT in a more elaborate way than Mehdad et al. (2010), and it does not depend on a specific MT approach. Another important difference between this work and that of Mehdad et al. (2011) is the set of features used for classification.

The paper is organized as follows: Section 2 describes the method used to collect the MT information and obtain the features; Section 3 explains the experimental framework; Section 4 shows the results obtained for the different features combination proposed; the paper ends with concluding remarks.

2 Features from machine translation

Our approach uses MT as a *black box* to detect parallelisms between the text fragments T and H by following these steps:

1. T is segmented in all possible sub-segments t_m^{m+p-1} of length p with $1 \leq p \leq L$ and $1 \leq m \leq |T| - p + 1$, where L is the maximum sub-segment length allowed. Analogously, H is segmented to get all the possible sub-segments h_n^{n+q-1} of length q, with $1 \leq q \leq L$ and $1 \leq n \leq |T| - q + 1$.
2. The sub-segments obtained from T are translated using all the available MT systems into L_H. Analogously, the sub-segments from H are translated into L_T, to generate a set of sub-segment pairs (t, h).
3. Those pairs of sub-segments (t, h) such that t is a sub-string of T and h is a sub-string of H are annotated as sub-segment links.

Note that it could be possible to use statistical MT to translate both T and H and then use word alignments to obtain the sub-segment links. However, we use this methodology to ensure that any kind of MT system can be used by our approach. As a result of this process, a sub-segment in T may be linked to more than one sub-segment in H, and vice versa. Based on these sub-segment links we have designed a set of features which may be used by a classifier for CLTE.

2.1 Basic features [Bas]

We used a set of basic features to represent the information from the sub-segment links between T and H, which are computed as the fraction of words in each of them covered by linked sub-segments of length $l \in [1, L]$. We define the feature function $F_l(S)$, applied on a text fragment S (either T or H) as:

$$F_l(S) = \mathrm{Cov}(S, l)/|S|$$

where $\mathrm{Cov}(S, l)$ is a function which obtains the number of words in S covered by at least one sub-segment of length l which is part of a sub-segment link. An additional feature is computed to represent the total proportion of words in each text fragment:

$$F_{\mathrm{total}}(S) = \mathrm{Cov}(S, *)/|S|$$

where $\mathrm{Cov}(S, *)$ is the same as $\mathrm{Cov}(S, l)$ but using sub-segments of any length up to L. $F_{\mathrm{total}}(S)$ provide information about overlapping that $F_l(S)$ cannot grasp. For instance, if we have $F_1(T) = 0.5$ and $F_2(T) = 0.5$, we cannot know if the sub-segments of $l = 1$ and $l = 2$ are covering the same or different words, so $F_{\mathrm{total}}(S)$ represents the actual proportion of words covered in a text fragment S. These feature functions are applied both on T and H, thus obtaining a set of $2 * L + 2$ features, henceforth Bas.

2.2 Extensions to the basic features

Some extensions can be made to the basic features defined above by using additional external resources. In this section we propose two extensions.

Separate analysis of function words and content words [Spl]. In this case, features represent, separately, function words, with poor lexical information, and content words, with richer lexical and semantic information. In this way, $F_l(S)$ is divided into $\mathrm{FF}_l(S)$ and $\mathrm{CF}_l(S)$ defined as:

$$\mathrm{FF}_l(S) = \mathrm{Cov}_{\mathrm{F}}(S, l)/|\mathrm{FW}(S)|$$

and

$$\mathrm{CF}_l(S) = \mathrm{Cov}_{\mathrm{C}}(S, l)/|\mathrm{CW}(S)|$$

where $\mathrm{FW}(S)$ is a function that returns the function words in text fragment S and $\mathrm{CW}(S)$ performs the same task for content words. Analogously, $\mathrm{Cov}_{\mathrm{F}}(S, l)$ and $\mathrm{Cov}_{\mathrm{C}}(S, l)$ are versions of $\mathrm{Cov}(S, l)$ which only consider function and content words, respectively. This extension can be also be applied to $F_{\mathrm{total}}(T)$ and $F_{\mathrm{total}}(H)$. The set of $4L + 4$ features obtained in this way (henceforth Spl) allows the classifier to use the information from the most relevant words in T and H to detect entailment.

Stemming [Stm and SplStm]. Stemming can also be used when detecting the sub-segment links. Both the table of sub-segment pairs and the text fragment pair (T,H) are stemmed before matching. In this way, conflicts of number or gender disagreement in the translations can be overcome in order to detect more sub-segment links. This new extension can be applied both to Bas, obtaining the set of features Stm, and to Spl, obtaining the set of features SplStm. Although lemmatization could have been used, stemming was preferred because it does not require the part-of-speech ambiguity to be solved, which may be difficult to solve when dealing with very short sub-segments.

2.3 Additional features

Two additional features were defined unrelated with the basic features proposed. The first one, called here R, is the length ratio $|T|/|H|$. Intuitively we can guess that if H is much longer than T it is unlikely that T entails H.

The second additional set of features is the one defined by Mehdad et al. (2011), so we will refer to it as M. The corresponding feature function computes, for the total number of sub-segments of a given length $l \in [1, L]$ obtained from a text fragment S, the fraction of them which appear in a sub-segment link. It is applied both to H and T and is defined as:

$$F'_l(S) = \mathrm{Linked}_l(S)/(|S| - l + 1)$$

where Linked_l is the number of sub-segments from S with length l which appear in a sub-segment link.

3 Experimental settings

The experiments designed for this task are aimed at evaluating the features proposed in Section 2. We evaluate our CLTE approach using the English–Spanish data sets provided in the task 8 of SemEval 2012 (Negri et al., 2012).

Datasets. Two datasets were provided by the organization of SemEval 2012 (Negri et al., 2011): a training set and a test set, both composed by a set of 500 pairs of sentences. CLTE detection is evaluated in both directions, so instances belong to one of these four classes: forward (the sentence in Spanish entails the one in English); backward (the sentence in English entails the one in Spanish); bidirectional (both sentences entail each other); and no entailment (neither of the sentences entails each other).

For the whole data set, both sentences in each instance were tokenized using the scripts[1] included in the Moses MT system (Koehn et al., 2007). Each sentence was segmented to get all possible sub-segments which were then translated into the other language.

External resources. We used three different MT systems to translate the sub-segments from English to Spanish, and vice versa:

- *Apertium*:[2] a free/open-source platform for the development of rule-based MT systems (Forcada et al., 2011). We used the English–Spanish MT system from the project's repository[3] (revision 34706).
- *Google Translate*:[4] an online MT system by Google Inc.
- *Microsoft Translator*:[5] an online MT system by Microsoft.

External resources were also used for the extended features described in Section 2.2. We used the stemmer[6] and the stopwords list provided by the SnowBall project for Spanish[7] and English.[8]

Classifier. We used the implementation of support vector machine included in the WEKA v.3.6.6 data mining software package (Hall et al., 2009) for multiclass classification, and a polynomial kernel.

4 Results and discussion

We tried the different features proposed in Section 2 in isolation, and also different combinations of them. Table 1 reports the accuracy for the different features described in Section 2 on the test set using sub-segments with lengths up to $L = 6$.[9]

[1] http://bit.ly/H4LNux
[2] http://www.apertium.org
[3] http://bit.ly/HCbn8a
[4] http://translate.google.com
[5] http://www.microsofttranslator.com
[6] http://bit.ly/H2HU97
[7] http://bit.ly/JMybmL
[8] http://bit.ly/Iwg9Vm
[9] All the results in this section are computed with $L = 6$, which proved to be the value providing the best accuracy for the dataset available after trying different values of L.

	Bas ∪ Spl ∪ Stm ∪ SplStm ∪ M ∪ R				Bas ∪ Spl ∪ M ∪ R			
	Apertium		**Ap.+Go.+Mi.**		**Apertium**		**Ap.+Go.+Mi.**	
	P	**R**	**P**	**R**	**P**	**R**	**P**	**R**
Backward	64.3%	64.8%	64.5%	72.8%	59.1%	64.8%	57.3%	60.0%
Forward	65.5%	57.6%	68.9 %	56.8%	59.8%	56.0%	58.7%	59.2%
Bidirectional	57.7%	56.8%	56.6%	55.2%	43.7%	41.6%	42.5%	40.8%
No-entailment	47.5%	53.6%	50.7%	54.4%	42.5%	43.2%	44.7%	44.0%
Accuracy	58.2%		59.8%		51.4%		51.0%	

Table 2: Precision (P) and recall (R) obtained by our approach for each of the four entailment classes and total accuracy on the English–Spanish test set using different feature combinations and different MT systems: Apertium, and a combination of Apertium, Google Translate, and Microsoft Translator (Ap.+Go.+Mi.).

Feature set	N_{f}	**Accuracy**
Bas	14	50.0%
Spl	28	56.0%
Stm	14	49.6%
SplStm	28	56.8%
R	1	45.8%
M	12	47.0%
Bas ∪ Spl	42	56.6%
Bas ∪ Stm	28	51.0%
Bas ∪ Spl ∪ Stm ∪ SplStm	84	57.4%
Bas ∪ Spl ∪ M ∪ R	41	58.2%
Bas ∪ Spl ∪ Stm ∪ ∪ SplStm ∪ M ∪ R	97	**59.8%**

Table 1: Accuracy obtained by the system using the different feature sets proposed in Section 2 for the test set. N_{f} is the number of features.

As can be seen, the features providing the best results on accuracy are the SplStm features. In addition, results show that all versions of the basic features (Bas, Spl, Stm, and SplStm) provide better results than the M feature alone. Some combinations of features are also reported in Table 1. Although many combinations were tried, we only report the results of the combinations of features performing best because of lack of space.

As can be seen, both feature combinations Bas ∪ Spl and Bas ∪ Stm obtain higher accuracy than the separated features. Combining all these features Bas ∪ Spl ∪ Stm ∪ SplStm provide even better results, thus confirming some degree of orthogonality between them. Combination Bas ∪ Spl ∪ M ∪ R obtains one of the best results, since it produces an improvement of almost 1% over combination Bas ∪ Spl ∪ Stm ∪ SplStm but using less than a half of features. Combining all the features provides the best accuracy as expected, so this seems to be the best combination for the task.

Table 2 reports the results sent for the SemEval 2012 task 8. We chose feature combinations Bas ∪ Spl ∪ M ∪ R and Bas ∪ Spl ∪ Stm ∪ SplStm ∪ M ∪ R since they are the best performing combinations. We sent two runs of our method using all three MT systems described in Section 3 and two more runs using only sub-segment translations from Apertium.

From the ten teams presenting systems for the contest, only one overcomes our best result. Even the results obtained using Apertium as the only MT system overcome seven of the ten approaches presented. This result confirms that state-of-the-art MT is a rich source of information for CLTE detection.

5 Concluding remarks

In this paper we have described a new method for CLTE detection which uses MT as a black-box source of bilingual information. We experimented with different features which were evaluated with the datasets for task 8 of SemEval 2012. We obtained up to 59.8% of accuracy on the Spanish–English test set provided, becoming the second best performing approach of the contest. As future works, we are now preparing experiments for other pairs of languages and we plan to use weights to promote those translations coming from more-reliable MT systems.

Acknowledgements: Work supported by the Spanish government through project TIN2009-14009-C02-01 and by Universitat d'Alacant through project GRE11-20. Google Translate service provided by the *University Research Program for Google Translate*. We thank M. Negri, Y. Mehdad, and M. Federico for encouraging us to participate in SemEval 2012.

References

Julio J. Castillo. 2011. A WordNet-based semantic approach to textual entailment and cross-lingual textual entailment. *International Journal of Machine Learning and Cybernetics*, 2(3):177–189.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer Berlin / Heidelberg.

Mikel Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1):10–18.

Sanda Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 905–912, Sydney, Australia.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Edmonton, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180, Prague, Czech Republic.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2010. Towards cross-lingual textual entailment. In *Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 321–324, Los Angeles, USA.

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*, pages 1336–1345, Portland, Oregon.

George A. Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. Source-language entailment modeling for translating unknown terms. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 791–799, Suntec, Singapore.

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*, pages 670–679, Edinburgh, United Kingdom.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa 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)*.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Chris Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 297–305, Suntec, Singapore.

Sergios Theodoridis and Konstantinos Koutroumbas. 2009. *Pattern Recognition*. Elsevier, 4th edition.