FBK: Cross-Lingual Textual Entailment Without Translation

Yashar Mehdad
FBK-irst
Trento , Italy
mehdad@fbk.eu

Matteo Negri
FBK-irst
Trento , Italy
negri@fbk.eu

José Guilherme C. de Souza
FBK-irst & University of Trento
Trento, Italy
desouza@fbk.eu

Abstract

This paper overviews FBK's participation in the Cross-Lingual Textual Entailment for Content Synchronization task organized within SemEval-2012. Our participation is characterized by using cross-lingual matching features extracted from lexical and semantic phrase tables and dependency relations. The features are used for multi-class and binary classification using SVMs. Using a combination of lexical, syntactic, and semantic features to create a cross-lingual textual entailment system, we report on experiments over the provided dataset. Our best run achieved an accuracy of 50.4% on the Spanish-English dataset (with the average score and the median system respectively achieving 40.7% and 34.6%), demonstrating the effectiveness of a "pure" cross-lingual approach that avoids intermediate translations.

1 Introduction

So far, cross-lingual textual entailment (CLTE) (Mehdad et al., 2010) has been applied to: *i)* available TE datasets (*"YES"/"NO"* uni-directional relations between monolingual pairs) transformed into their cross-lingual counterpart by translating the hypotheses into other languages (Negri and Mehdad, 2010), and *ii)* machine translation evaluation datasets (Mehdad et al., 2012b). The content synchronization task represents a challenging application scenario to test the capabilities of CLTE systems, by proposing a richer inventory of phenomena (i.e. *"Bidirectional"/"Forward"/"Backward"/"No entailment"* multi-directional entailment relations).

Multi-directional CLTE recognition can be seen as the identification of semantic equivalence and information disparity between two topically related sentences, at the cross-lingual level. This is a core aspect of the multilingual content synchronization task, which represents a challenging application scenario for a variety of NLP technologies, and a shared research framework for the integration of semantics and MT technology.

The CLTE methods proposed so far adopt either a "pivoting approach" (translation of the two input texts into the same language, as in (Mehdad et al., 2010)), or an "integrated solution" that exploits bilingual phrase tables to capture lexical relations and contextual information (Mehdad et al., 2011). The promising results achieved with the integrated approach still rely on phrasal matching techniques that disregard relevant semantic aspects of the problem. By filling this gap integrating linguistically motivated features, in our participation, we propose an approach that combines lexical, syntactic and semantic features within a machine learning framework (Mehdad et al., 2012a).

Our submitted runs have been produced by training and optimizing multiclass and binary SVM classifiers, over the Spanish-English (Spa-Eng) development set. In both cases, our results were positive, showing significant improvements over the median systems and average scores obtained by participants. The overall results confirm the difficulty of the task, and the potential of our approach in combining linguistically motivated features in a "pure" cross-lingual approach that avoids the recourse to external MT components.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 701–705,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

2 Experiments

In our experiment we used the Spa-Eng portion of the dataset described in (Negri et al., 2012; Negri et al., 2011), consisting of 500 multi-directional entailment pairs which was provided to train the systems and 500 pairs for the submission. Each pair in the dataset is annotated with “Bidirectional”, “Forward”, “Backward” or “No entailment” judgements.

2.1 Approach

Our system builds on the integration of lexical, syntactic and semantic features in a supervised learning framework. Our model builds on three main feature sets, respectively derived from: *i)* phrase tables, *ii)* dependency relations, and *iii)* semantic phrase tables.

1. Phrase Table (PT) matching: through these features, a semantic judgement about entailment is made exclusively on the basis of lexical evidence. The matching features are calculated with a phrase-to-phrase matching process. A phrase in our approach is an *n-gram* composed of one or more (up to 5) consecutive words, excluding punctuation. Entailment decisions are assigned combining phrasal matching scores calculated for each level of *n-grams* (*i.e.* considering the number of *1-grams*, *2-grams*,..., *5-grams* extracted from H that match with *n-grams* in T). Phrasal matches, performed either at the level of tokens, lemmas, or stems, can be of two types:

1. **Exact**: in the case that two phrases are identical at one of the three levels (token, lemma, stem).
2. **Lexical**: in the case that two different phrases can be mapped through entries of the resources used to bridge T and H (*i.e.* phrase tables).

For each phrase in H, we first search for exact matches at the level of token with phrases in T. If no match is found at a token level, the other levels (lemma and stem) are attempted. Then, in case of failure with exact matching, lexical matching is performed at the same three levels. To reduce redundant matches, the lexical matches between pairs of phrases which have already been identified as exact matches are not considered.

Once the matching phase for each *n-gram* level has been concluded, the number of matches $Match_n$ and the number of phrases in the hypothesis $H(n)$ is used to estimate the portion of phrases in H that are matched at each level *n* (Equation 1).[1] Since languages can express the same meaning with different amounts of words, a phrase with length *n* in H can match a phrase with any length in T.

$$Match_n = \frac{Match_n}{|H(n)|} \tag{1}$$

In order to build English-Spanish phrase tables for our experiments, we used the freely available Europarl V.4, News Commentary and United Nations Spanish-English parallel corpora released for the WMT10 Shared Translation Task.[2] We run the TreeTagger (Schmid, 1995) and Snowball stemmer (Porter, 2001) for preprocessing, and used the Giza++ (Och and Ney, 2000) toolkit to align the tokenized corpora at the word level. Subsequently, we extracted the bi-lingual phrase table from the aligned corpora using the Moses toolkit (Koehn et al., 2007).

2. Dependency Relation (DR) matching targets the increase of CLTE precision. By adding syntactic constraints to the matching process, DR features aim to reduce wrong matches often occurring at the lexical level. For instance, the contradiction between “*Yahoo acquired Overture*” and “*Overture compró Yahoo*” is evident when syntax (in this case subject-object inversion) is taken into account, but can not be caught by bag-of-words methods.

We define a dependency relation as a triple that connects pairs of words through a grammatical relation. For example, *“nsubj (loves, John)”* is a dependency relation with head *loves* and dependent *John* connected by the relation *nsubj*, which means that *“John”* is the *subject* of *“loves”*. DR matching captures similarities between dependency relations, by combining the syntactic and lexical level. In a valid match, while the relation has to be the same (“exact”

[1] When checking for entailment from H to T, the normalization is carried out dividing the number of n-grams in H by the number of n-grams in T. The same holds for dependency relation and semantic phrase table matching.

[2] http://www.statmt.org/wmt10/

match), the connected words must be either the same or semantically equivalent in the two languages. For example, *"nsubj (loves, John)"* can match *"nsubj (ama, John)"* and *"nsubj (quiere, John)"* but not *"dobj (quiere, John)"*.

Given the dependency tree representations of T and H, for each grammatical relation (*r*) we calculate a DR matching score ($Match_r$, see Equation 2) as the number of matching occurrences of *r* in T and H (respectively $DR_r(T)$ and $DR_r(H)$), divided by the number of occurrences of *r* in H.

$$match_r = \frac{|match(DR_r(T), DR_r(H))|}{|DR_r(H)|} \qquad (2)$$

In our experiments, in order to extract dependency relation (DR) matching features, the dependency tree representations of English and Spanish texts have been produced with DepPattern (Otero and Lopez, 2011). We then mapped the sets of dependency relation labels for the English-Spanish parser output into: Adjunct, Determiner, Object, Subject and Preposition. The dictionary, containing about 9M bilingual word pairs, created during the alignment of the English-Spanish parallel corpora provided the lexical knowledge to perform matches when the connected words are different.

3. Semantic Phrase Table (SPT) matching: represents a novel way to leverage the integration of semantics and MT-derived techniques. To this aim, SPT improves CLTE methods relying on pure lexical match, by means of "generalized" phrase tables annotated with shallow semantic labels. Semantically enhanced phrase tables, with entries in the form "*[LABEL]* $word_1...word_n$ *[LABEL]*" (*e.g.* "*[ORG] acquired [ORG]*"), are used as a recall-oriented complement to the lexical phrase tables used in machine translation (token-based entries like "*Yahoo acquired Overture*"). The main motivation for this augmentation is that word replacement with semantic tags allows to match T-H tokens that do not occur in the original bilingual parallel corpora used for phrase table extraction. Our hypothesis is that the increase in recall obtained from relaxed matches through semantic tags in place of "out of vocabulary" terms (*e.g.* unseen person, location, or organization names) is an effective way to improve CLTE performance, even at the cost of some loss in precision. Semantic phrase tables, however, have two additional advantages. The first is related to their smaller size and, in turn, its positive impact on system's efficiency, due to the considerable search space reduction. Semantic tags allow to merge different sequences of tokens into a single tag and, consequently, different phrase entries can be unified to one semantic phrase entry. As a result, for instance, the SPT used in our experiments is more than 30% smaller than the original token-based one. The second advantage relates to their potential impact on the confidence of CLTE judgements. Since a semantic tag might cover more than one token in the original entry phrase, SPT entries are often short generalizations of longer original phrases. Consequently, the matching process can benefit from the increased probability of mapping higher order n-grams (*i.e.* those providing more contextual information) from H into T and vice-versa.

Like lexical phrase tables, SPTs are extracted from parallel corpora. As a first step, we annotate the corpora with named-entity taggers (FreeLing in our case (Carreras et al., 2004)) for the source and target languages, replacing named entities with general semantic labels chosen from a coarse-grained taxonomy including the categories: person, location, organization, date and numeric expression. Then, we combine the sequences of unique labels into one single token of the same label, and we run Giza++ (Och and Ney, 2000) to align the resulting semantically augmented corpora. Finally, we extract the semantic phrase table from the augmented aligned corpora using the Moses toolkit (Koehn et al., 2007).

For the matching phase, we first annotate T and H in the same way we labeled our parallel corpora. Then, for each n-gram order (n=1 to 5, excluding punctuation), we use the SPT to calculate a matching score (SPT_match_n, see Equation 3), as the number of n-grams in H that match with phrases in T divided by the number of n-grams in H. The matching algorithm is same as the phrase table matching one.

$$SPT\_match_n = \frac{|SPT_n(H) \cap SPT(T)|}{|SPT_n(H)|} \qquad (3)$$

Run	Features	Classification	Parameter selection	Result
1	PT+SPT+DR	Multiclass	Entire training set	0.502
2	PT+SPT+DR	Multiclass	2-fold cross validation	0.490
3	PT+SPT+DR	Binary	Entire training set	**0.504**
4	PT+SPT+DR	Binary	2-fold cross validation	0.500

Table 1: Summary of the submitted runs and results for Spa-Eng dataset.

Forward			**Backward**			**No entailment**			**Bidirectional**		
P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	F1
0.515	**0.704**	0.595	0.546	0.568	0.557	0.447	0.304	0.362	0.482	0.440	0.460

Table 2: Best run's Precision/Recall/F1 scores.

In our supervised learning framework, the computed PT, SPT and DR scores are used as separate features, giving to an SVM classifier, LIBSVM (Chang and Lin, 2011), the possibility to learn optimal feature weights from training data.

2.2 Submitted runs

In order to test our models under different conditions, we set the CLTE problem both as two-way and multiclass classification tasks.

Two-way classification casts multidirectional entailment as a unidirectional problem, where each pair is analyzed checking for entailment both from left to right and from right to left. In this condition, each original test example is correctly classified if both pairs originated from it are correctly judged ("YES-YES" for bidirectional, "YES-NO" for forward, "NO-YES" for backward entailment, and "NO-NO" for no entailment). Two-way classification represents an intuitive solution to capture multidirectional entailment relations but, at the same time, a suboptimal approach in terms of efficiency since two checks are performed for each pair.

Multiclass classification is more efficient, but at the same time more challenging due to the higher difficulty of multiclass learning, especially with small datasets. We also tried to use the parameter selection tool for C-SVM classification using the RBF (radial basis function) kernel, available in LIBSVM package. Our submitted runs and results have been obtained with the settings summarized in table 1.

As can be seen from the table, our best result has been achieved by Run 3 (50.4% accuracy), which is significantly higher than the average and median score over the best runs obtained by participants (44.0% and 40.7% respectively). The detailed results achieved by the best run are reported in Table 2. We can observe that our system is performing well for recognizing the unidirectional entailment (i.e. forward and backward), while the performance drops over no_entailment pairs. The low results for bidirectional cases also reflect the difficulty of discriminating the no_entailment pairs from the bidirectional ones. Looking at the detailed results, we can observe a high recall in the forward and backward entailment cases, which could be explained by the effectiveness of the semantic phrase table matching features aiming at coverage increase over lexical methods. Adding more linguistically motivated features and weighting the non-matched phrases can be a starting point to improve the overall results for other cases (bidirectional and no entailment).

3 Conclusion

In this paper we described our participation to the cross-lingual textual entailment for content synchronization task at SemEval-2012. We approached this task by combining lexical, syntactic and semantic features, at the cross-lingual level without recourse to intermediate translation steps. In spite of the difficulty and novelty of the task, our results on the Spanish-English dataset (0.504) prove the effectiveness of the approach with significant improvements over the reported average and median accuracy scores for the 29 submitted runs (respectively 40.7% and 34.6%).

Acknowledgments

This work has been partially supported by the EU-funded project CoSyne (FP7-ICT-4-248531).

References

X. Carreras, I. Chao, L. Padró, and M. Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

C.C. Chang and C.J. Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3).

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL 2007)*.

Y. Mehdad, M. Negri, and M. Federico. 2010. Towards Cross-Lingual Textual Entailment. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*.

Y. Mehdad, M. Negri, and M. 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 (ACL HLT 2011)*.

Y. Mehdad, M. Negri, and M. Federico. 2012a. Detecting Semantic Equivalence and Information Disparity in Cross-lingual Documents. In *Proceedings of the ACL'12*.

Y. Mehdad, M. Negri, and M. Federico. 2012b. Match without a Referee: Evaluating MT Adequacy without Reference Translations. In *Proceedings of the Machine Translation Workshop (WMT2012)*.

M. Negri and Y. Mehdad. 2010. Creating a Bi-lingual Entailment Corpus through Translations with Mechanical Turk: $100 for a 10-day rush. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 212–216. Association for Computational Linguistics.

M. Negri, L. Bentivogli, Y. Mehdad, D. Giampiccolo, and A. Marchetti. 2011. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of EMNLP 2011*.

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)*.

F.J. Och and H. Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*.

P.G. Otero and I.G. Lopez. 2011. A Grammatical Formalism Based on Patterns of Part-of-Speech Tags. *International journal of corpus linguistics*, 16(1).

M. Porter. 2001. Snowball: A language for stemming algorithms.

H. Schmid. 1995. Treetaggera language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*.