Detecting Semantic Equivalence and Information Disparity in Cross-lingual Documents

Yashar Mehdad **Matteo Negri** **Marcello Federico**
Fondazione Bruno Kessler, FBK-irst
Trento , Italy
{mehdad|negri|federico}@fbk.eu

Abstract

We address a core aspect of the multilingual content synchronization task: the identification of novel, more informative or semantically equivalent pieces of information in two documents about the same topic. This can be seen as an application-oriented variant of textual entailment recognition where: *i)* T and H are in different languages, and *ii)* entailment relations between T and H have to be checked in both directions. Using a combination of lexical, syntactic, and semantic features to train a cross-lingual textual entailment system, we report promising results on different datasets.

1 Introduction

Given two documents about the same topic written in different languages (*e.g.* Wiki pages), content synchronization deals with the problem of automatically detecting and resolving differences in the information they provide, in order to produce aligned, mutually enriched versions. A roadmap towards the solution of this problem has to take into account, among the many sub-tasks, the identification of information in one page that is semantically equivalent, novel, or more informative with respect to the content of the other page. In this paper we set such problem as an application-oriented, cross-lingual variant of the Textual Entailment (TE) recognition task (Dagan and Glickman, 2004). Along this direction, we make two main contributions:

(a) Experiments with multi-directional cross-lingual textual entailment. So far, cross-lingual textual entailment (CLTE) has been only applied to: *i)* available TE datasets (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 (MT) evaluation datasets (Mehdad et al., 2012). Instead, we experiment with the only corpus representative of the multilingual content synchronization scenario, and the richer inventory of phenomena arising from it (multi-directional entailment relations).

(b) Improvement of current CLTE methods. The CLTE methods proposed so far adopt either a "*pivoting approach*" based on the translation of the two input texts into the same language (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, however, still rely on phrasal matching techniques that disregard relevant semantic aspects of the problem. By filling this gap integrating linguistically motivated features, we propose a novel approach that improves the state-of-the-art in CLTE.

2 CLTE-based content synchronization

CLTE has been proposed by (Mehdad et al., 2010) as an extension of textual entailment which consists of deciding, given a text T and an hypothesis H *in different languages*, if the meaning of H can be inferred from the meaning of T. The adoption of entailment-based techniques to address content synchronization looks promising, as several issues inherent to such task can be formalized as entailment-related prob-

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 120–124,
Jeju, Republic of Korea, 8-14 July 2012.

lems. Given two pages (*P1* and *P2*), these issues include identifying, and properly managing:
(1) Text portions in *P1* and *P2* that express the same meaning (bi-directional entailment). In such cases no information has to migrate across *P1* and *P2*, and the two text portions will remain the same;
(2) Text portions in *P1* that are more informative than portions in *P2* (forward entailment). In such cases, the entailing (more informative) portions from *P1* have to be translated and migrated to *P2* in order to replace or complement the entailed (less informative) fragments;
(3) Text portions in *P2* that are more informative than portions in *P1* (backward entailment), and should be translated to replace or complement them;
(4) Text portions in *P1* describing facts that are not present in *P2*, and vice-versa (the "unknown" cases in RTE parlance). In such cases, the novel information from both sides has to be translated and migrated in order to mutually enrich the two pages;
(5) Meaning discrepancies between text portions in the two pages ("contradictions" in RTE parlance).

CLTE has been previously modeled as a phrase matching problem that exploits dictionaries and phrase tables extracted from bilingual parallel corpora to determine the number of word sequences in H that can be mapped to word sequences in T. In this way a *semantic* judgement about entailment is made exclusively on the basis of *lexical* evidence. When only unidirectional entailment relations from T to H have to be determined (RTE-like setting), the full mapping of the hypothesis into the text usually provides enough evidence for a positive entailment judgement. Unfortunately, when dealing with multi-directional entailment, the correlation between the proportion of matching terms and the correct entailment decisions is less strong. In such framework, for instance, the full mapping of the hypothesis into the text is *per se* not sufficient to discriminate between forward entailment and semantic equivalence. To cope with these issues, we explore the contribution of syntactic and semantic features as a complement to lexical ones in a supervised learning framework.

3 Beyond lexical CLTE

In order to enrich the feature space beyond pure lexical match through phrase table entries, our model builds on two additional feature sets, derived from *i)* semantic phrase tables, and *ii)* dependency relations.

Semantic Phrase Table (SPT) matching represents a novel way to leverage the integration of semantics and MT-derived techniques. SPT matching extends CLTE methods based on pure lexical match by means of "generalized" phrase tables annotated with shallow semantic labels. SPTs, with entries in the form "*[LABEL]* $word_1 ... word_n$ *[LABEL]*", are used as a recall-oriented complement to the phrase tables used in MT. A motivation for this augmentation is that semantic tags allow to match 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 names) is an effective way to improve CLTE performance, even at the cost of some loss in precision.

Like lexical phrase tables, SPTs are extracted from parallel corpora. As a first step we annotate the parallel corpora with named-entity taggers for the source and target languages, replacing named entities with general semantic labels chosen from a coarse-grained taxonomy (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) we use the SPT to calculate a matching score as the number of n-grams in H that match with phrases in T divided by the number of n-grams in H.[1]

Dependency Relation (DR) matching targets the increase of CLTE precision. Adding syntactic constraints to the matching process, DR features aim to reduce the amount of wrong matches often occurring with bag-of-words methods (both at the lexical level and with recall-oriented SPTs). For instance, the contradiction between "*Yahoo acquired*

[1]When checking for entailment from H to T, the normalization is carried out dividing by the number of n-grams in T.

Overture" and "*Overture compró Yahoo*", which is evident when syntax is taken into account, can not be caught by shallow methods. We define a dependency relation as a triple that connects pairs of words through a grammatical relation. DR matching captures similarities between dependency relations, combining the syntactic and lexical level. In a valid match, while the relation has to be the same, the connected words can be either the same, or semantically equivalent terms in the two languages (*e.g.* according to a bilingual dictionary). Given the dependency tree representations of T and H, for each grammatical relation (*r*) we calculate a DR matching score as the number of matching occurrences of *r* in T and H, divided by the number of occurrences of *r* in H. Separate DR matching scores are calculated for each relation *r* appearing both in T and H.

4 Experiments and results

4.1 Content synchronization scenario

In our first experiment we used the English-German portion of the CLTE corpus described in (Negri et al., 2011), consisting of 500 multi-directional entailment pairs which we equally divided into training and test sets. Each pair in the dataset is annotated with "Bidirectional", "Forward", or "Backward" entailment judgements. Although highly relevant for the content synchronization task, "Contradiction" and "Unknown" cases (*i.e.* "NO" entailment in both directions) are not present in the annotation. However, this is the only available dataset suitable to gather insights about the viability of our approach to multi-directional CLTE recognition.[2] We chose the ENG-GER portion of the dataset since for such language pair MT systems performance is often lower, making the adoption of simpler solutions based on pivoting more vulnerable.

To build the English-German phrase tables we combined the Europarl, News Commentary and "denews"[3] parallel corpora. After tokenization, Giza++ and Moses were respectively used to align the corpora and extract a lexical phrase table (PT). Similarly, the semantic phrase table (SPT) has been extracted from the same corpora annotated with the Stanford NE tagger (Faruqui and Padó, 2010; Finkel et al., 2005). Dependency relations (DR) have been extracted running the Stanford parser (Rafferty and Manning, 2008; De Marneffe et al., 2006). The dictionary created during the alignment of the parallel corpora provided the lexical knowledge to perform matches when the connected words are different, but semantically equivalent in the two languages. To combine and weight features at different levels we used SVMlight (Joachims, 1999) with default parameters.

In order to experiment under testing conditions of increasing complexity, we set the CLTE problem both as a two-way and as a three-way classification task. Two-way classification casts multi-directional 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, and "NO-YES" for backward 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. Three-way classification is more efficient, but at the same time more challenging due to the higher difficulty of multiclass learning, especially with small datasets.

Results are compared with two pivoting approaches, checking for entailment between the original English texts and the translated German hypotheses.[4] The first (Pivot-EDITS), uses an optimized distance-based model implemented in the open source RTE system EDITS (Kouylekov and Negri, 2010; Kouylekov et al., 2011). The second (Pivot-PPT) exploits paraphrase tables for phrase matching, and represents the best monolingual model presented in (Mehdad et al., 2011). Table 1 demonstrates the success of our results in proving the two main claims of this paper. **(a)** In both settings all the feature sets used outperform the approaches taken as terms of comparison. The 61.6% accuracy achieved in the most challenging setting

[2]Recently, a new dataset including "Unknown" pairs has been used in the "*Cross-Lingual Textual Entailment for Content Synchronization*" task at SemEval-2012 (Negri et al., 2012).

[3]http://homepages.inf.ed.ac.uk/pkoehn/

[4]Using Google Translate.

	PT	PT+DR	PT+SPT	PT+SPT+DR	Pivot-EDITS	Pivot-PPT
Cont. Synch. (2-way)	57.8	58.6	62.4	**63.3**	27.4	57.0
Cont. Synch. (3-way)	57.4	57.8	58.7	**61.6**	25.3	56.1
					RTE-3 AVG	Pivot PPT
RTE3-derived	62.6	63.6	63.5	**64.5**	62.4	63.5

Table 1: CLTE accuracy results over content synchronization and RTE3-derived datasets.

(3-way) demonstrates the effectiveness of our approach to capture meaning equivalence and information disparity in cross-lingual texts.
(b) In both settings the combination of lexical, syntactic and semantic features (PT+SPT+DR) significantly improves[5] the state-of-the-art CLTE model (PT). Such improvement is motivated by the joint contribution of SPTs (matching more and longer n-grams, with a consequent recall improvement), and DR matching (adding constraints, with a consequent gain in precision). However, the performance increase brought by DR features over PT is minimal. This might be due to the fact that both PT and DR features are precision-oriented, and their effectiveness becomes evident only in combination with recall-oriented features (SPT).

Cross-lingual models also significantly outperform pivoting methods. This suggests that the noise introduced by incorrect translations makes the pivoting approach less attractive in comparison with the more robust cross-lingual models.

4.2 RTE-like CLTE scenario

Our second experiment aims at verifying the effectiveness of the improved model over RTE-derived CLTE data. To this aim, we compare the results obtained by the new CLTE model with those reported in (Mehdad et al., 2011), calculated over an English-Spanish entailment corpus derived from the RTE-3 dataset (Negri and Mehdad, 2010).

In order to build the English-Spanish lexical phrase table (PT), we used the Europarl, News Commentary and United Nations parallel corpora. The semantic phrase table (SPT) was extracted from the same corpora annotated with FreeLing (Carreras et al., 2004). Dependency relations (DR) have been extracted parsing English texts and Spanish hypotheses with DepPattern (Gamallo and Gonzalez, 2011).

[5] $p < 0.05$, calculated using the approximate randomization test implemented in (Padó, 2006).

Accuracy results have been calculated over 800 test pairs of the CLTE corpus, after training the SVM binary classifier over the 800 development pairs. Our new features have been compared with: *i)* the state-of-the-art CLTE model (PT), *ii)* the best monolingual model (Pivot-PPT) presented in (Mehdad et al., 2011), and *iii)* the average result achieved by participants in the monolingual English RTE-3 evaluation campaign (RTE-3 AVG). As shown in Table 1, the combined feature set (PT+SPT+DR) significantly[5] outperforms the lexical model (64.5% vs 62.6%), while SPT and DR features separately added to PT (PT+SPT, and PT+DR) lead to marginal improvements over the results achieved by the PT model alone (about 1%). This confirms the conclusions drawn from the previous experiment, that precision-oriented and recall-oriented features lead to a larger improvement when they are used in combination.

5 Conclusion

We addressed the identification of semantic equivalence and information disparity in two documents about the same topic, written in different languages. 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. Casting the problem as a CLTE task, we extended previous lexical models with syntactic and semantic features. Our results in different cross-lingual settings prove the feasibility of the approach, with significant state-of-the-art improvements also on RTE-derived data.

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 Language Resources and Evaluation Conference (LREC 2004)*, volume 4.

I. Dagan and O. Glickman. 2004. Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. In *Proceedings of the PASCAL Workshop of Learning Methods for Text Understanding and Mining*.

M.C. De Marneffe, B. MacCartney, and C.D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC 2006)*, volume 6, pages 449–454.

M. Faruqui and S. Padó. 2010. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of the 10th Conference on Natural Language Processing (KONVENS 2010)*, Saarbrücken, Germany.

J.R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 2005)*.

P. Gamallo and I. Gonzalez. 2011. A grammatical formalism based on patterns of part of speech tags. *International Journal of Corpus Linguistics*, 16(1):45–71.

T. Joachims. 1999. Advances in kernel methods. chapter Making large-scale support vector machine learning practical, pages 169–184. MIT Press, Cambridge, MA, USA.

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 on Association for Computational Linguistics, Demonstration Session (ACL 2007)*.

M. Kouylekov and M. Negri. 2010. An Open-Source Package for Recognizing Textual Entailment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, system demonstrations (ACL 2010)*.

M. Kouylekov, Y. Mehdad, and M. Negri. 2011. Is it Worth Submitting this Run? Assess your RTE System with a Good Sparring Partner. *Proceedings of the EMNLP TextInfer 2011 Workshop on Textual Entailment*.

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. 2012. 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 Amazonś Mechanical Turk*.

M. Negri, L. Bentivogli, Y. Mehdad, D. Giampiccolo, and A. Marchetti. 2011. Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textual Entailment Corpora. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (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)*.

S. Padó, 2006. *User's guide to* `sigf`*: Significance testing by approximate randomisation*.

A.N. Rafferty and C.D. Manning. 2008. Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In *In Proceedings of the ACL 2008 Workshop on Parsing German*.