

Towards Cross-Lingual Textual Entailment

Yashar Mehdad^{1,2}, Matteo Negri¹, Marcello Federico¹

FBK-Irst¹, University of Trento²

Trento, Italy

{mehdad, negri, federico}@fbk.eu

Abstract

This paper investigates *cross-lingual textual entailment* as a semantic relation between two text portions in different languages, and proposes a prospective research direction. We argue that cross-lingual textual entailment (CLTE) can be a core technology for several cross-lingual NLP applications and tasks. Through preliminary experiments, we aim at proving the feasibility of the task, and providing a reliable baseline. We also introduce new applications for CLTE that will be explored in future work.

1 Introduction

Textual Entailment (TE) (Dagan and Glickman, 2004) has been proposed as a generic framework for modeling language variability. Given two texts T and H, the task consists in deciding if the meaning of H can be inferred from the meaning of T. So far, TE has been only applied in a *monolingual* setting, where both texts are assumed to be written in the same language. In this work, we propose and investigate a *cross-lingual* extension of TE, where we assume that T and H are written in different languages.

The great potential of integrating (monolingual) TE recognition components into NLP architectures has been reported in several works, such as question answering (Harabagiu and Hickl, 2006), information retrieval (Clinchant et al., 2006), information extraction (Romano et al., 2006), and document summarization (Lloret et al., 2008).

To the best of our knowledge, mainly due to the absence of cross-lingual TE (CLTE) recognition

components, similar improvements have not been achieved yet in any cross-lingual application. As a matter of fact, despite the great deal of attention that TE has received in recent years (also witnessed by five editions of the Recognizing Textual Entailment Challenge¹), interest for cross-lingual extensions has not been in the mainstream of TE research, which until now has been mainly focused on the English language.

Nevertheless, the strong interest towards cross-lingual NLP applications (both from the market and research perspectives, as demonstrated by successful evaluation campaigns such as CLEF²) is, to our view, a good reason to start investigating CLTE, as well. Along such direction, research can now benefit from recent advances in other fields, especially machine translation (MT), and the availability of: *i*) large amounts of parallel and comparable corpora in many languages, *ii*) open source software to compute word-alignments from parallel corpora, and *iii*) open source software to set-up strong MT baseline systems. We strongly believe that all these resources can potentially help in developing inference mechanisms on multilingual data.

Building on these considerations, this paper aims to put the basis for future research on the cross-lingual Textual Entailment task, in order to allow for semantic inference across languages in different NLP tasks. Among these, as a long-term goal, we plan to adopt CLTE to support the alignment of text portions that express the same meaning in different languages. As a possible application scenario, CLTE

¹<http://pascallin.ecs.soton.ac.uk/Challenges/RTE/>

²www.clef-campaign.org/

can be used to address *content merging* tasks in tidy multilingual environments, such as commercial Web sites, digital libraries, or user generated content collections. Within such framework, as it will be discussed in the last section of this paper, CLTE components can be used for automatic content synchronization in a concurrent, collaborative, and multilingual editing setting, *e.g.* Wikipedia.

2 Cross Lingual Textual Entailment

Adapting the definition of TE we define CLTE as a relation between two natural language portions in different languages, namely a text T (*e.g.* in English), and a hypothesis H (*e.g.* in French), that holds if a human after reading T would infer that H is most likely true, or otherwise stated, the meaning of H can be entailed (inferred) from T .

We can see two main orthogonal directions for approaching CLTE: *i)* simply bring CLTE back to the monolingual case by translating H into the language of T , or vice-versa; *ii)* try to embed cross-lingual processing techniques inside the TE recognition process. In the following, we briefly overview and motivate each approach.

Basic approaches. The simplest approach is to add a MT component to the front-end of an existing TE engine. For instance, let the French hypothesis H be translated into English and then run the TE engine on T and the translation of H . There are several good reasons to follow this divide-and-conquer approach, as well as some drawbacks. Decoupling the cross-lingual and the entailment components results in a simple and modular architecture that, according to well known software engineering principles, results easier to develop, debug, and maintain. Moreover, a decoupled CLTE architecture would allow for easy extensions to other languages as it just requires extra MT systems. Along the same idea of pivoting through English, in fact, the same TE system can be employed to perform CLTE between any language pair, once MT is available from each language into English. A drawback of the decoupled approach is that as MT is still far from being perfect, translation errors are propagated to the TE engine and might likely affect performance. To cope with this issue, we explored the alternative approach of applying TE on a list of n -best translations provided

by the MT engine, and take a final decision based on some system combination criterion. This latter approach potentially reduces the impact of translation errors, but might significantly increase the computational requirements of CLTE.

Advanced approaches. The idea is to move towards a cross-lingual TE approach that takes advantage of a tighter integration of MT and TE algorithms and techniques. This could result in methods for recognizing TE across languages without translating the texts and, in principle, with a lower complexity. When dealing with phrase-based statistical MT (Koehn et al., 2007), a possible approach is to extract information from the phrase-table to enrich the inference and entailment rules which could be used in a distance based entailment system. As an example the entailment relations between the French phrase “*ordinateur portable*” and the English phrase “*laptop*”, or between the German phrase “*europaeischen union*” and the English word “*Europe*” could be captured from parallel corpora through statistical phrase-based MT approaches.

There are several implications that make this approach interesting. First of all, we believe that research on CLTE can employ inference mechanisms and semantic knowledge sources to augment existing MT methods, leading to improvements in the translation quality (*e.g.* (Padó et al., 2009)). In addition, the acquired rules could as well enrich the available multilingual resources and dictionaries such as MultiWordNet³.

3 Feasibility studies

The main purpose of our preliminary experiments is to verify the feasibility of CLTE, as well as setting baseline results to be further improved over time. To this aim, we started by adopting the basic approach previously discussed. In particular, starting from an English/French corpus of T-H pairs, we automatically translated each H fragment from French into English.

Our decisions build on several motivations. First of all, the reason for setting English and French as a first language pair for experiments is to rely on higher quality translation models, and larger amounts of parallel data for future improvements.

³<http://multiwordnet.fbk.eu/>

Second, the reason for translating the hypotheses is that, according to the notion of TE, they are usually shorter, less detailed, and barely complex in terms of syntax and concepts with respect to the texts. This makes them easier to translate preserving the original meaning. Finally, from an application-oriented perspective, working with English Ts seems more promising due the richness of English data available (*e.g.* in terms of language variability, and more detailed elaboration of concepts). This increases the probability to discover entailment relations with Hs in other languages.

In order to create a realistic and standard setting, we took advantage of the available RTE data, selecting the RTE3 development set and manually translating the hypotheses into French. Since the manual translation requires trained translators, and due to time and logistics constraints, we obtained 520 translated hypotheses (randomly selected from the entire RTE3 development set) which built our bilingual entailment corpus for evaluation.

In the initial step, following our basic approach, we translated the French hypotheses to English using Google⁴ and Moses⁵. We trained a phrase-base translation model using Europarl⁶ and News Commentary parallel corpora in Moses, applying a 6-gram language model trained on the New York Times portion of the English Gigaword corpus⁷.

As a TE engine, we used the EDITS⁸ package (Edit Distance Textual Entailment Suite). This system is an open source software package based on edit distance algorithms, which computes the T-H distance as the cost of the edit operations (*i.e.* insertion, deletion and substitution) that are necessary to transform T into H. By defining the edit distance algorithm and a cost scheme (*i.e.* which defines the costs of each edit operation), this package is able to learn a distance model over a set of training pairs, which is used to decide if an entailment relation holds over each test pair.

In order to obtain a monolingual TE model, we trained and tuned (Mehdad, 2009) our model on the RTE3 test set, to reduce the overfitting bias, since

our original data was created over the RTE3 development set. Moreover, we used a set of lexical entailment rules extracted from Wikipedia and WordNet, as described in (Mehdad et al., 2009). To begin with, we used this model to classify the created cross-lingual entailment corpus in three different settings: 1) hypotheses translated by Google, 2) hypotheses translated by Moses (1st best), and 3) the original RTE3 monolingual English pairs.

Results reported in Table 1 show that using Google as a translator, in comparison with the original manually-created data, does not cause any drop in performance. This confirms that merely translating the hypothesis using a very good translation model (Google) is a feasible and promising direction for CLTE. Knowing that Google has one of the best French-English translation models, the downward trend of results using Moses translator, in contrast with Google, is not out of our expectation. Trying to bridge this gap brings us to the next round of experiments, where we extracted the n -best trans-

| | Orig. | Google | Moses 1st best | Moses 30 best | Moses > 0.4 |
|-------------|--------------|---------------|--------------------------|-------------------------|-----------------------|
| Acc. | 63.48 | 63.48 | 61.37 | 62.90 | 62.90 |

Table 1: Results comparison over 520 test pairs.

lations produced by Moses, to have a richer lexical variability, beneficial for improving the TE recognition. The graph in Figure 1 shows an incremental improvement when the n -best translated hypotheses are used. Besides that, trying to reach a more monotonic distribution of the results, we normalized the ranking score (from 0 to 1) given by Moses, and in each step we chose the first n results over a normalized score. In this way, having the hypotheses with the score of above 0.4, we achieved the highest accuracy of 62.9%. This is exactly equal to adopting the 30-best hypotheses translated by Moses. Using this method, we could improve the performance up to 1.5% above the 1st best results, achieving almost the same level of performance obtained with Google.

4 A possible application scenario

Among the many possible applications, the task of managing textual information in multiple languages represents an ideal application scenario for CLTE. Along such direction, our long-term goal is to use

⁴<http://translate.google.com>

⁵<http://www.statmt.org/moses/>

⁶<http://www.statmt.org/europarl/>

⁷<http://www ldc.upenn.edu>

⁸<http://edits.fbk.eu/>

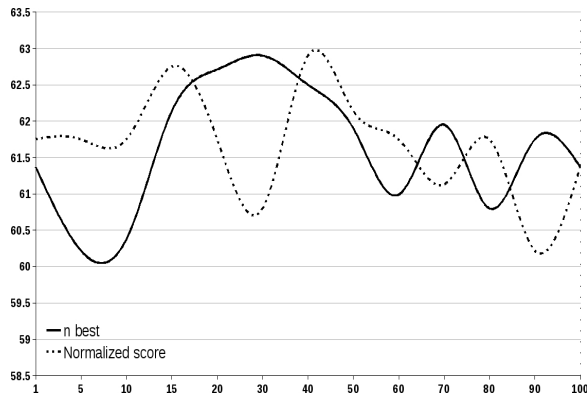


Figure 1: Accuracy gained by n -best Moses translations.

CLTE components in the task of *synchronizing* the content of documents about the same topic (e.g. Wikipedia articles), written in different languages. Currently, multilingual Wikis rely on users to manually translate different Wiki pages on the same subject. This is not only a time-consuming procedure but also the source of many inconsistencies, as users update the different language versions separately, and every update would require translators to compare the different language versions and synchronize the updates. Our goal is to automate this process by integrating MT and CLTE in a two-step process where: *i*) CLTE is used to identify text portions that should “migrate” from one page to the other, and *ii*) MT is used to actually translate these portions in the appropriate target language.

The adoption of entailment-based techniques to address the multilingual content synchronization task looks promising, as several issues inherent to such task can be formalized as TE-related problems. Given two pages (P1 and P2), these issues include identifying (and then properly managing):

1. Text portions in P1 and P2 that express exactly the same meaning (bi-directional entailment, or semantic equivalence) and which should not migrate across pages;

2. Text portions in P1 that are more specific than portions of P2 (unidirectional entailment between P2 and P1 or vice-versa) and should replace them;

3. Text portions in P1 describing facts that are not present in P2, and which should be added in P2 or vice-versa (the “unknown” cases in RTE parlance);

4. Meaning discrepancies between text portions in P1 and text portions in P2 (“contradictions” in

RTE parlance).

5 Conclusion

This paper presented a preliminary investigation towards cross-lingual Textual Entailment, focusing on possible research directions and alternative methodologies. Baseline results have been provided to demonstrate the potentialities of a simple approach that integrates MT and monolingual TE components. Overall, our work sets a novel framework for further studies and experiments to improve cross-lingual NLP tasks. In particular, CLTE can be scaled to more complex problems, such as cross-lingual content merging and synchronization.

Acknowledgments

This work has been partially supported by the EC-funded project CoSyne (FP7-ICT-4-24853)

References

- S. Clinchant, C. Goutte, and E. Gaussier. 2006. Lexical entailment for information retrieval. In *Proc. ECIR'06*.
- I. Dagan and O. Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. *Proc. of the PASCAL Workshop of Learning Methods for Text Understanding and Mining*.
- S. Harabagiu and A. Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proc. COLING/ACL 2006*.
- P. Koehn et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL07 Demo and Poster Sessions*.
- E. Lloret, Ó. Ferrández, R. Muñoz, and M. Palomar. 2008. A text summarization approach under the influence of textual entailment. In *Proc. NLPCS 2008*.
- Y. Mehdad, M. Negri, E. Cabrio, M. Kouylekov, and B. Magnini. 2009. Edits: An open source framework for recognizing textual entailment. In *Proc. TAC 2009. To appear*.
- Yashar Mehdad. 2009. Automatic cost estimation for tree edit distance using particle swarm optimization. In *Proc. ACL '09*.
- S. Padó, M. Galley, D. Jurafsky, and C. D. Manning. 2009. Textual entailment features for machine translation evaluation. In *Proc. StatMT '09*.
- L. Romano, M. Kouylekov, I. Szpektor, I. Dagan, and A. Lavelli. 2006. Investigating a generic paraphrase-based approach for relation extraction. In *Proc. EACL 2006*.