ICT: A Translation based Method for Cross-lingual Textual Entailment

Fandong Meng, Hao Xiong and Qun Liu
Key Lab. of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{mengfandong,xionghao,liuqun}@ict.ac.cn

Abstract

In this paper, we present our system description in task of Cross-lingual Textual Entailment. The goal of this task is to detect entailment relations between two sentences written in different languages. To accomplish this goal, we first translate sentences written in foreign languages into English. Then, we use EDITS[1], an open source package, to recognize entailment relations. Since EDITS only draws monodirectional relations while the task requires bidirectional prediction, thus we exchange the hypothesis and test to detect entailment in another direction. Experimental results show that our method achieves promising results but not perfect results compared to other participants.

1 Introduction

In Cross-Lingual Textual Entailment task (CLTE) of 2012, the organizers hold a task for Cross-Lingual Textual Entailment. The Cross-Lingual Textual Entailment task addresses textual entailment (TE) recognition under a new dimension (cross-linguality), and within a new challenging application scenario (content synchronization) Readers can refer to M. Negri et al. 2012.s., for more detailed introduction. [1]

Textual entailment, on the other hand, recognize, generate, or extract pairs of natural language expressions, and infer that if one element is true, whether the other element is also true. Several methods are proposed by previous researchers. There have been some workshops on textual entailment in recent years. The recognizing textual entailment challenges (Bar-Haim et al. 2006; Giampiccolo, Magnini, Dagan, & Dolan, 2007; Giampiccolo, Dang, Magnini, Dagan, & Dolan, 2008), currently in the 7th year, provide additional significant thrust. Consequently, there are a large number of published articles, proposed methods, and resources related to textual entailment. A special issue on textual entailment was also recently published, and its editorial provides a brief overview of textual entailment methods (Dagan, Dolan, Magnini, & Roth, 2009).

Textual entailment recognizers judge whether or not two given language expressions constitute a correct textual entailment pair. Different methods may operate at different levels of representation of the input expressions. For example, they may treat the input expressions simply as surface strings, they may operate on syntactic or semantic representations of the input expressions, or on representations combining information from different

[1] http://edits.fbk.eu/

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 715–720,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

levels. Logic-based approach is to map the language expressions to logical meaning representations, and then rely on logical entailment checks, possibly by invoking theorem provers (Rinaldi et al., 2003; Bos & Markert, 2005; Tatu & Moldovan, 2005, 2007). An alternative to use logical meaning representations is to start by mapping each word of the input language expressions to a vector that shows how strongly the word co-occurs with particular other words in corpora (Lin, 1998b), possibly also taking into account syntactic information, for example requiring that the co-occurring words participate in particular syntactic dependencies (Pad´o & Lapata, 2007). Several textual entailment recognizing methods operate directly on the input surface strings. For example, they compute the string edit distance (Levenshtein, 1966) of the two input strings, the number of their common words, or combinations of several string similarity measures (Malakasiotis & Androutsopoulos, 2007). Dependency grammar parsers (Melcuk, 1987; Kubler, McDonald, & Nivre, 2009) are popular in textual entailment research. However, cross-lingual textual entailment brings some problems on past algorithms. On the other hand, many methods can't be applied to it directly.

In this paper, we propose a translation based method for cross-lingual textual entailment, which has been described in Mehdad et al. 2010. First, we translate one part of the text, which termed as "t1" and written in one language, into English, which termed as "t2". Then, we use EDITS, an open source package, to recognize entailment relations between two parts. Large-scale experiments are conducted on four language pairs, French-English, Spanish-English, Italian-English and German-English. Although our method achieves promising results reported by organizers, it is still far from perfect compared to other participants.

The remainder of this paper is organized as follows. We describe our system framework in section 2. We report experimental results in section 3 and draw our conclusions in the last section.

2 System Description

Figure 1 illustrates the overall framework of our system, where a machine translation model is employed to translate foreign language into English, since original EDITS could only deal with the text in the same language pairs.

In the following of this section, we will describe the translation module and configuration of EDITS in details.

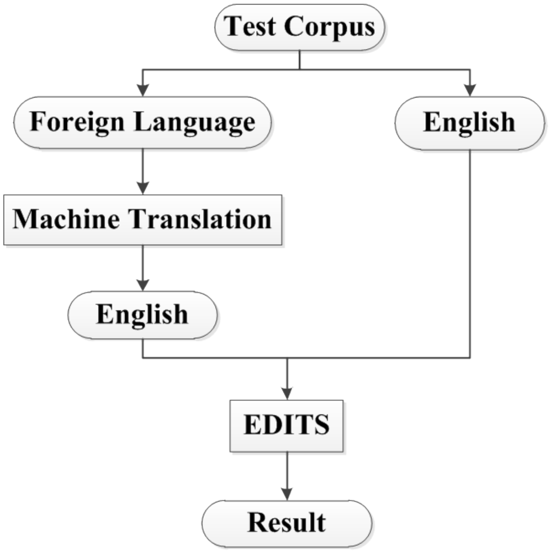


Figure 1: The framework of our system.

2.1 Machine Translation

Recently, machine translation has attracted intensive attention and has been well studied in natural language community. Effective models, such as Phrase-Based model (Koehn et al., 2003), Hierarchical Phrase-Based model (HPB) (Chiang, 2005), and Syntax-Based (Liu et al., 2006) model have been proposed to improve the translation quality. However, since current translation models require parallel corpus to extract translation rules, while parallel corpus on some language pairs such as Italian-English and Spanish-English are hard to obtain, therefore, we could use Google Translation Toolkit (GTT) to generate translation.

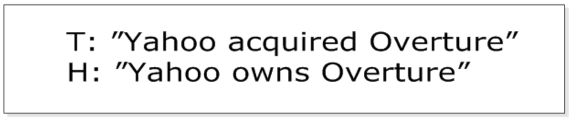
Specifically, WMT[2] released some bilingual corpus for training, thus we use some portion to train a French-English translation engine using hierarchical phrase-based model. We also exploit system combination technique (A Rosti et al., 2007) to improve translation quality via blending the translation of our models and GTT's. It is worth noting that GTT only gives 1-best translation, thus we duplicate 50 times to generate 50-best for system combination.

[2] http://www.statmt.org/wmt12/

2.2 Textual Entailment

Many methods have been proposed to recognize textual entailment relations between two expressions written in the same language. Since edit distance algorithms are effective on this task, we choose this method. And we use popular toolkit, EDITS, to accomplish the textual entailment task.

EDITS is an open source software, which is used for recognizing entailment relations between two parts of text, termed as “T” and “H”. The system is based on the edit distance algorithms, and 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”. EDITS requires that three modules are defined: an edit distance algorithm, a cost scheme for the three edit operations, and a set of rules expressing either entailment or contradiction. Each module can be easily configured by the user as well as the system parameters. EDITS can work at different levels of complexity, depending on the linguistic analysis carried on over “T” and “H”. Both linguistic processors and semantic resources that are available to the user can be integrated within EDITS, resulting in a flexible, modular and extensible approach to textual entailment.

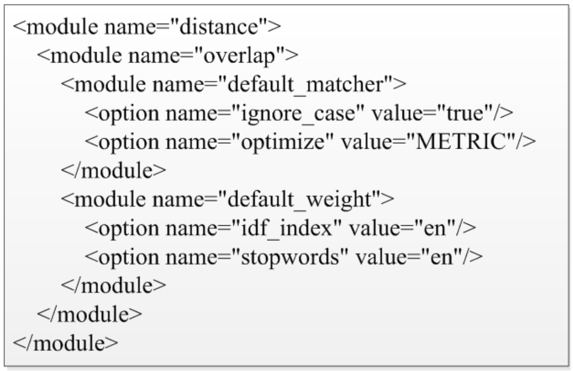


T: ”Yahoo acquired Overture”
H: ”Yahoo owns Overture”

Figure 2: An Example of two expressions EDITS can recognize.

Figure 2 shows an example of two expressions that EDITS can recognize. EDITS will give an answer that whether expression “H” is true given that expression “T” is true. The result is a Boolean value. If “H” is true given “T” is true, then the result is “YES”, otherwise “NO”.

EDITS implements a distance-based framework which assumes that the probability of an entailment relation between a given “T”-“H” pair is inversely proportional to the distance between “T” and “H” (i.e. the higher the distance, the lower is the probability of entailment). Within this framework the system implements and harmonizes different approaches to distance computation, providing both edit distance algorithms, and similarity algorithms. Each algorithm returns a normalized distance score (a number between 0 and 1). At a training stage, distance scores calculated over annotated “T”-“H” pairs are used to estimate a threshold that best separates positive from negative examples. The threshold, which is stored in a Model, is used at a test stage to assign an entailment judgment and a confidence score to each test pair.



```
<module name="distance">
  <module name="overlap">
    <module name="default_matcher">
      <option name="ignore_case" value="true"/>
      <option name="optimize" value="METRIC"/>
    </module>
    <module name="default_weight">
      <option name="idf_index" value="en"/>
      <option name="stopwords" value="en"/>
    </module>
  </module>
</module>
```

Figure 3: Our configured file for training

Figure 3 shows our configuration file for training models, we choose “distance” algorithm in EDITS, and “default_matcher”, and “ignore_case” , and some other default but effective configured parameters.

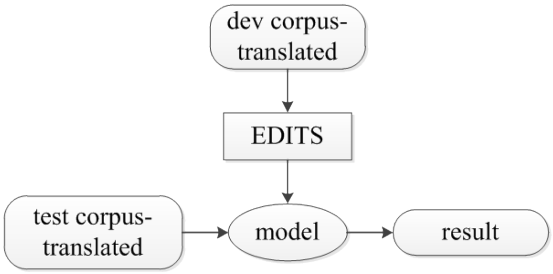


Figure 4: The overall training and decoding procedure in our system.

Figure 4 shows our training and decoding procedure. As EDITS can only recognize textual entailment from one part to the other, we manually change the tag “H” with “T”, and generate the results again, and then compute two parts’ entailment relations. For example, if “T”-“H” is “YES”, and “H”-“T” is “NO”, then the entailment result between them is “forward”; if “T”-“H” is “NO”, and “H”-“T” is “YES”, then the entailment result between them is “backward”; if both “T”-“H” and “H”-“T” are “YES”, the result is “bidirectional”;

otherwise “no_entailment”.

3 Experiments and Results

Since organizers of SemEval 2012 task 8 supply a piece of data for training, we thus exploit it to optimize parameters for EDITS. Table 1 shows the F-measure score of training set analyzed by EDITS, where “FE” represents French-English, “SE” represents Spanish-English, “IE” represents Italian-English and “GE” represents Italian-English.

Judgment	**FE**	**SE**	**IE**	**GE**
forward	0.339	0.373	0.440	0.327
backward	0.611	0.574	0.493	0.552
no_entailment	0.533	0.535	0.494	0.494
bidirectional	0.515	0.502	0.506	0.495
Overall	**0.516**	**0.506**	**0.488**	**0.482**

Table 1: Results on training set.

From Table 1, we can see that the performance of “forward” prediction is lower than others. One explanation is that the “T” is translated from foreign language, which is error unavoidable. Thus some rules used for checking “T”, such as stop-word list will be disabled. Then it is possible to induce a “NO” relation between “T” and “H” that results in lower recall of “forward”.

Since for French-English, we build a system combination for improving the quality of translation. Table 2 shows the results of BLEU score of translation quality, and F-score of entailment judgment.

System	**BLEU4**	**F-score**
HPB	28.74	0.496
GTT	30.08	0.508
COMB	**30.57**	**0.516**

Table 2: Performance of different translation model, where COMB represents system combination.

From table 2, we find that the translation quality slightly affect the correctness of entailment judgment. However, the difference of performance in entailment judgment is smaller than that in translation quality. We explain that the translation models exploit phrase-based rules to direct the translation, and the translation errors mainly come from the disorder between each phrases. While a distance based entailment model generally considers the similarity of phrases between test and hypothesis, thus the disorder of phrases influences the judgment slightly.

Using the given training data for tuning parameters, table 3 to table 6 shows the detailed experimental results on testing data, where P represents precision and R indicates recall, and both of them are calculated by given evaluation script.

French -- English			
Judgment	**P**	**R**	**F-measure**
forward	0.750	0.192	0.306
backward	0.517	0.496	0.506
no_entailment	0.385	0.656	0.485
bidirectional	0.444	0.480	0.462
Overall	0.456		
Best System	0.570		

Table 3: Test results on French-English

Spanish -- English			
Judgment	**P**	**R**	**F-measure**
forward	0.750	0.240	0.364
backward	0.440	0.472	0.456
no_entailment	0.395	0.560	0.464
bidirectional	0.436	0.520	0.474
Overall	0.448		
Best System	0.632		

Table 4: Test results on Spanish-English

Italian – English			
Judgment	**P**	**R**	**F-measure**
forward	0.661	0.296	0.409
backward	0.554	0.368	0.442
no_entailment	0.427	0.448	0.438
bidirectional	0.383	0.704	0.496
Overall	0.454		
Best System	0.566		

Table 5: Test results on Italian-English

German – English			
Judgment	**P**	**R**	**F-measure**
forward	0.718	0.224	0.341
backward	0.493	0.552	0.521
no_entailment	0.390	0.512	0.443
bidirectional	0.439	0.552	0.489
Overall	0.460		
Best System	0.558		

Table 6: Test results on German-English

After given golden testing reference, we also investigate the effect of training set to testing set. We choose testing set from RTE1 and RTE2, both are English text, as our training set for optimization of EDITS, and the overall results are shown in table 7 to table 10, where CLTE is training set given by this year's organizers.

French -- English			
Judgment	**CLTE**	**RTE1**	**RTE2**
forward	0.306	0.248	0.289
backward	0.506	0.425	0.440
no_entailment	0.485	0.481	0.485
bidirectional	0.462	**0.472**	**0.485**
Overall	0.456	0.430	0.444

Table 7: Test results on French-English given different training set.

Spanish – English			
Judgment	**CLTE**	**RTE1**	**RTE2**
forward	0.364	0.293	0.297
backward	0.456	0.332	0.372
no_entailment	0.464	0.386	0.427
bidirectional	0.474	**0.484**	**0.503**
Overall	0.448	0.400	0.424

Table 8: Test results on Spanish-English given different training set.

Italian -- English			
Judgment	**CLTE**	**RTE1**	**RTE2**
forward	0.409	0.333	0.335
backward	0.442	0.394	0.436
no_entailment	0.438	0.410	0.421
bidirectional	0.496	0.474	0.480
Overall	0.454	0.420	0.432

Table 9: Test results on Italian-English given different training set.

German – English			
Judgment	**CLTE**	**RTE1**	**RTE2**
forward	0.341	**0.377**	**0.425**
backward	0.521	0.372	0.460
no_entailment	0.443	0.437	**0.457**
bidirectional	0.489	0.487	**0.508**
Overall	0.460	0.434	**0.470**

Table 10: Test results on German-English given different training set.

Results in table 7 and table 8 shows that models trained on "CLTE" have better performance than those trained on RTE1 and RTE2, except "bidirectional" judgment type. In Table 9, all results decoding by models trained on "CLTE" are the best. And in Table 10, only a few results decoding by models trained on "RTE1" and "RTE2" have higher score. The reason may be that, the test corpora are bilingual, there are some errors in the machine translation procedure when translate one part of the test from its language into the other. When training on these bilingual text and decoding these bilingual text, these two procedure have error consistency. Some errors may be counteracted. If we train on RTE, a standard monolingual text, and decode a bilingual text, more errors may exist between the two procedures. So we believe that, if we use translation based strategy (machine translation and monolingual textual entailment) to generate cross-lingual textual entailment, we should use translation based strategy to train models, rather than use standard monolingual texts.

4 Conclusion

In this paper, we demonstrate our system framework for this year's cross-lingual textual entailment task. We propose a translation based model to address cross-lingual entailment. We first translate all foreign languages into English, and then employ EDITS to induce entailment relations. Experiments show that our method achieves promising results but not perfect results compared to other participants.

Acknowledgments

The authors were supported by National Science Foundation of China, Contracts 90920004, and High-Technology R&D Program (863) Project No 2011AA01A207 and 2012BAH39B03. We thank organizers for their generous supplied resources and arduous preparation. We also thank anonymous reviewers for their thoughtful suggestions.

References

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., & Szpektor, I. 2006.*The 2nd PASCAL recognising textual entailment challenge*. In Proc. of the 2nd PASCAL ChallengesWorkshop on Recognising Textual Entailment, Venice, Italy.

Bos, J., & Markert, K. 2005. *Recognising textual entailment with logical inference.* In Proc. Of the Conf. on HLT and EMNLP, pp. 628–635, Vancouver, BC, Canada.

Dagan, I., Dolan, B., Magnini, B., & Roth, D. 2009. Recognizing textual entailment: Rational,evaluation and approaches. Nat. Lang. Engineering, 15(4), i–xvii. Editorial of the special issue on Textual Entailment.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In Proceedings of ACL 2005, pages 263–270.

Giampiccolo, D., Dang, H., Magnini, B., Dagan, I., & Dolan, B. 2008. *The fourth PASCAL recognizing textual entailment challenge.* In Proc. of the Text Analysis Conference, pp. 1–9, Gaithersburg, MD.

Giampiccolo, D., Magnini, B., Dagan, I., & Dolan, B. 2007. *The third PASCAL recognizing textual entailment challenge.* In Proc. of the ACL-Pascal Workshop on Textual Entailment and Paraphrasing, pp. 1–9, Prague, Czech Republic.

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

Ion Androutsopoulos and Prodromos Malakasiotis. 2010.*A Survey of Paraphrasing and Textual Entailment Methids.* Journal of Artificial Intelligence Research, 32, 135-187.

Kouylekov, M. and Negri, M. 2010. *An open-source package for recognizing textual entailment.* Proceedings of the ACL 2010 System Demonstrations, 42-47.

Kubler, S., McDonald, R., & Nivre, J. 2009. *Dependency Parsing. Synthesis Lectures on HLT.* Morgan and Claypool Publishers.

Levenshtein, V. 1966. *Binary codes capable of correcting deletions, insertions, and reversals.* Soviet Physice-Doklady, 10, 707–710.

Lin, D. 1998b. *An information-theoretic definition of similarity.* In Proc. of the 15th Int. Conf. on Machine Learning, pp. 296–304, Madison, WI. Morgan Kaufmann, San Francisco, CA.

Malakasiotis, P., & Androutsopoulos, I. 2007. *Learning textual entailment using SVMs and string similarity measures.* In Proc. of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 42–47, Prague. ACL.

Mehdad, Y. and Negri, M. and Federico, M.2010. *Towards Cross-Lingual Textual Entailment. Human Language Technologies.*The 2010 Annual Conference of the NAACL. 321-324.

Mehdad, Y. and Negri, M. and Federico, M.2011. *Using bilingual parallel corpora for cross-lingual textual entailment*. Proceedings of ACL-HLT

Melcuk, I. 1987. *Dependency Syntax: Theory and Practice.* State University of New York Press.

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

Negri, M. and Bentivogli, L. and Mehdad, Y. and Giampiccolo, D. and Marchetti, A.2011. *Divide and conquer: crowdsourcing the creation of cross-lingual textual entailment corpora.* Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Pad´o, S., & Lapata, M. 2007. *Dependency-based construction of semantic space models.* Comp. Ling., 33(2), 161–199.

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

Rinaldi, F., Dowdall, J., Kaljurand, K., Hess, M., & Molla, D. 2003. *Exploiting paraphrases in a question answering system.* In Proc. of the 2nd Int. Workshop in Paraphrasing, pp. 25–32, Saporo, Japan.

Rosti, A. and Matsoukas, S. and Schwartz, R. *Improved word-level system combination for machine translation,* ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS,2007

Tatu, M., & Moldovan, D. 2005. *A semantic approach to recognizing textual entailment.* In Proc. of the Conf. on HLT and EMNLP, pp. 371–378, Vancouver, Canada.

Tatu, M., & Moldovan, D. 2007. *COGEX at RTE 3.* In Proc. of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 22–27, Prague, Czech Republic.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree–to string alignment template for statistical machine translation. In Proceedings of ACL 2006, pages 609–616, Sydney, Australia, July.