ECNUCS: Recognizing Cross-lingual Textual Entailment Using Multiple Text Similarity and Text Difference Measures

Jiang ZHAO
Department of Computer
Science and Technology
East China Normal University
Shanghai, P.R.China
51121201042@ecnu.cn

Man LAN[*]
Department of Computer
Science and Technology
East China Normal University
Shanghai, P.R.China
mlan@cs.ecnu.edu.cn

Zheng-Yu NIU
Baidu Inc.
Beijing, P.R.China
niuzhengyu@baidu.com

Abstract

This paper presents our approach used for cross-lingual textual entailment task (task 8) organized within SemEval 2013. Cross-lingual textual entailment (CLTE) tries to detect the entailment relationship between two text fragments in different languages. We solved this problem in three steps. Firstly, we use a off-the-shelf machine translation (MT) tool to convert the two input texts into the same language. Then after performing a text preprocessing, we extract multiple feature types with respect to surface text and grammar. We also propose novel feature types regarding to sentence difference and semantic similarity based on our observations in the preliminary experiments. Finally, we adopt a multiclass SVM algorithm for classification. The results on the cross-lingual data collections provided by SemEval 2013 show that (1) we can build portable and effective systems across languages using MT and multiple effective features; (2) our systems achieve the best results among the participants on two test datasets, i.e., FRA-ENG and DEU-ENG.

1 Introduction

The Cross-lingual Textual Entailment (CLTE) task in SemEval 2013 consists in detecting the entailment relationship between two topic-related text fragments (usually called **T**(ext) and **H**(ypothesis)) in different languages, which is a cross-lingual extension of TE task in (Dagan and Glickman, 2004). We say T entails H if the meaning of H can be inferred from the meaning of T. Mehdad et al. (2010b) firstly proposed this problem within a new challenging application scenario, i.e., content synchronization. In consideration of the directionality, the task needs to assign one of the following entailment judgments to a pair of sentences (1) forward: unidirectional entailment from T to H; (2) backward: unidirectional entailment from H to T; (3) bidirectional: the two fragments entail each other (i.e., semantic equivalence); (4) non-entailment: there is no entailment between T and H.

During the last decades, many researchers and communities have paid a lot of attention to resolve the TE detection (e.g., seven times of the Recognizing Textual Entailment Challenge, i.e., from RTE1 to RET7, have been held) since identifying the relationship between two sentences is at the core of many NLP applications, such as text summarization (Lloret et al., 2008) or question answering (Harabagiu and Hickl, 2006). For example, in text summarization, a redundant sentence should be omitted from the summary if this sentence can be entailed from other expressions in the summary. CLTE extends those tasks with lingual dimensionality, where more than one language is involved. Although it is a relatively new task, a basic solution has been provided in (Mehdad et al., 2010b), which brings the problem back to monolingual scenario using MT to translate H into the language of T. The promising performance indicates the potentialities of such a simple approach which integrates MT and monolingual TE algorithms (Castillo, 2011; Jimenez et al., 2012; Mehdad et al., 2010a).

In this work, we regard CLTE as a multiclass classification problem, in which multiple feature types are used in conjunction with a multiclass SVM classifier. Specifically, our approach can be divided into three steps. Firstly, following (Esplà-Gomis et al., 2012; Meng et al., 2012), we use MT to

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 118–123, Atlanta, Georgia, June 14-15, 2013.

bridge the gap of language differences between T and H. Secondly, we perform a preprocessing procedure to maximize the similarity of the two text fragments so as to make a more accurate calculation of surface text similarity measures. Besides several features described in previous work (Malakasiotis, 2009; Esplà-Gomis et al., 2012), we also propose several novel features regarding to sentence difference and semantic similarity. Finally, all these features are combined together and serves as input of a multiclass SVM classifier. After analyzing of the results obtained in preliminary experiments, we also cast this problem as a hierarchical classification problem.

The remainder of the paper is organized as follows. Section 2 describes different features used in our systems. Section 3 presents the system settings including the datasets and preprocessing. Section 4 shows the results of different systems on different language pairs. Finally, we conclude this paper with future work in Section 5.

2 Features

In this section, we will describe a variety of feature types used in our experiments.

2.1 Basic features

The BC feature set consists of length measures on variety sets including $|A|, |B|, |A-B|, |B-A|, |A \cup B|, |A \cap B|, |A|/|B|$ and $|B|/|A|$, where A and B represent two texts, and the length of set is the number of non-repeated elements in this set. Once we view the text as a set of words, $A - B$ means the set of words found in A but not in B, $A \cup B$ means the set of words found in either A or B and $A \cap B$ means the set of shared words found in both A and B.

Given a pair of texts, i.e., <T,H>, which are in different languages, we use MT to translate one of them to make them in the same language. Thus, we can get two pairs of texts, i.e., <T^t,H> and <T,H^t>. We apply the above eight length measures to the two pairs, resulting in a total of 16 features.

2.2 Surface Text Similarity features

Following (Malakasiotis and Androutsopoulos, 2007), the surface text similarity (STS) feature set contains nine similarity measures:

Jaccard coefficient: It is defined as $\frac{|A \cap B|}{|A \cup B|}$, where $|A \cap B|$ and $|A \cup B|$ are as in the BC.

Dice coefficient: Defined as $\frac{2*|A \cap B|}{|A|+|B|}$.

Overlap coefficient: This is the following quantity, $Overlap(A, B) = \frac{|A \cap B|}{|A|}$.

Weighted overlap coefficient: We assign the *tf*idf* value to each word in the sentence to distinguish the importance of different words. The weighted overlap coefficient is defined as follows:

$$WOverlap(A, B) = \frac{\sum_{w_i \in A \cap B} W_{w_i}}{\sum_{w_i \in A} W_{w_i}},$$

where W_{w_i} is the weight of word w_i.

Cosine similarity: $\cos(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{\|\overrightarrow{x}\| \cdot \|\overrightarrow{y}\|}$, where $\overrightarrow{x}$ and $\overrightarrow{y}$ are vectorial representations of texts (i.e. A and B) in $tf * idf$ schema.

Manhattan distance: Defined as $M(\overrightarrow{x}, \overrightarrow{y}) = \sum\limits_{i=1}^{n} |x_i - y_i|$.

Euclidean distance: Defined as $E(\overrightarrow{x}, \overrightarrow{y}) = \sqrt{\sum\limits_{i=1}^{n} (x_i - y_i)^2}$.

Edit distance: This is the minimum number of operations needed to transform A to B. We define an operation as an insertion, deletion or substitution of a word.

Jaro-Winker distance: Following (Winkler and others, 1999), the Jaro-Winkler distance is a measure of similarity between two strings at the word level.

In total, we can get 11 features in this feature set.

2.3 Sematic Similarity features

Almost every previous work used the surface texts or exploited the meanings of words in the dictionary to calculate the similarity of two sentences rather than the actual meaning in the sentence. In this feature set (SS), we introduce a latent model to model the semantic representations of sentences since latent models are capable of capturing the contextual meaning of words in sentences. We used weighted textual matrix factorization (WTMF) (Guo and Diab, 2012) to model the semantics of the sentences. The model factorizes the original term-sentence matrix X into two matrices such that $X_{i,j} \approx P_{*,i}^T Q_{*,j}$, where $P_{*,i}$ is a latent semantics vector profile for word w_i and $Q_{*,j}$ is the vector profile that represents the sentence s_j. The weight matrix W is introduced in the optimization process in order to model the missing words at the right level of emphasis. We propose three similarity measures according to different strategies:

wtw: word-to-word based similarity defined as $sim(A, B) = \lg \frac{\sum_{w_i \in A} W_{w_i} \cdot \max_{w_j \in B}(P_{*,i}, P_{*,j})}{\sum_{w_i \in A} W_{w_i}}$.

wts: word-to-sentence based similarity defined as $sim(A, B) = \lg \frac{\sum_{w_i \in A} W_{w_i} \cdot P_{*,i} \cdot Q_{*,k}}{\sum_{w_i \in A} W_{w_i}}$.

sts: sentence-to-sentence based similarity defined as $sim(A, B) = \lg (Q_{*,i} \cdot Q_{*,j})$.

Also we calculate the cosine similarity, Euclidean and Manhattan distance, weighted overlap coefficient using those semantics vectors, resulting in 10 features.

2.4 Sentence Difference features

Most of those above measures are symmetric and only a few are asymmetric, which means they may not be very suitable for the task that requires dealing with directional problems. We solve this problem by introducing sentence difference measures.

We observed that many entailment relationships between two sentences are determined by only tiny parts of the sentences. As a result, the similarity of such two sentences by using above measures will be close to 1, which may mislead the classifier. Furthermore, almost all similarity measures in STS are symmetric, which means the same similarity has no help to distinguish the different directions. Based on the above considerations, we propose a novel sentence difference (SD) feature set to discover the differences between two sentences and tell the classifier the possibility the entailment should not hold.

The sentence difference features are extracted as follows. Firstly, a word in one sentence is considered as matched if we can find the same word in the other sentence. Then we find all matched words and count the number of unmatched words in each sentence, resulting in 2 features. If one sentence has no unmatched words, we say that this sentence can be entailed by the other sentence. That is, we can infer the entailment class through the number of unmatched words. We regard this label as our third feature type. Secondly, different POS types of unmatched words may have different impacts on the classification, therefore we count the number of unmatched words in each sentence that belong to a small set of POS tags (here consider only NN, JJ, RB, VB and CD tags), which produces 10 features, resulting in a total of 13 sentence difference features.

2.5 Grammatical Relationship features

The grammatical relationship feature type (GR) is designed to capture the grammatical relationship between two sentences. We first replace the words in a sentence with their part-of-speech (POS) tags, then apply the STS measures on this new “sentence”.

In addition, we use the Stanford Parser to get the dependency information represented in a form of relation units (e.g. nsubj(example, this)). We calculate the BC measures on those units and the overlap coefficients together with the harmonic mean of them. Finally, we get 22 features.

2.6 Bias features

The bias features (BS) are to check the differences between two sentences in certain special aspects, such as polarity and named entity. We use a method based on subjectivity of lexicons (Loughran and McDonald, 2011) to get the polarity of a sentence by simply comparing the numbers of positive and negative words. If the numbers are the same, then we set the feature to 1, otherwise -1. Also, we check whether one sentence entails the other using only the named entity information. We consider four categories of named entities, i.e., person, organization, location, number, which are recognized by using the Stanford NER toolkit. We set the feature to 1 if the named entities in one sentence are found in the other sentence, otherwise -1. As a result, this feature set contains 9 features.

3 Experimental Setting

We evaluated our approach using the data sets provided in the task 8 of SemEval 2013 (Negri et al., 2013). The data sets consist of a collection of 1500 text fragment pairs (1000 for training consisting of training and test set in SemEval 2012 and 500 for test) in each language pair. Four different language pairs are provided: German-English, French-English, Italian-English and Spanish-English. See (Negri et al., 2013) for more detailed description.

3.1 Preprocess

We performed the following text preprocessing. Firstly, we employed the state-of-the-art Statistical Machine Translator, i.e., Google translator, to translate each pair of texts <T,H> into <T^t,H> and <T,H^t>, thus they were in the same language. Then we extracted all above described feature sets from the pair <T^t,H> (note that <T,H^t> are also used in BC), so the below steps were mainly operated on this pair. After that, all sentences were tokenized and lemmatized using the Stanford Lemmatizer and all stop words were removed, followed by the equivalent replacement procedure. The replacement procedure consists of the following 3 steps:

Abbreviative replacement. Many phrases or organizations can be abbreviated to a set of capitalized letters, e.g. *"New Jersey"* is usually wrote as *"NJ"* for short. In this step, we checked every word whose length is 2 or 3 and if it is the same as the "word" consisting of the first letters of the successive words in another sentence, then we replaced it by them.

Semantic replacement. We observed that although some lemmas in H and T were in the different forms, they actually shared the same meaning, e.g. *"happen"* and *"occur"*. Here, we focused on replacing a lemma in one sentence with another lemma in the other sentence if they were: 1) in the same synonymy set; or 2) gloss-related. Two lemmas were gloss-related if a lemma appeared in the gloss of the other. For example, the gloss of "trip" is *"a journey for some purpose"* (WordNet 2.1 was used for looking up the synonymy and gloss of a lemma), so the lemma "journey" is gloss-related with "trip". No word sense disambiguation was performed and all synsets for a particular lemma were considered.

Context replacement. The context of a lemma is defined as the non-stopword lemmas around it. Given two text fragments, i.e., **T.** *...be erroneously label as a "register sex offender."* and **H.** *...be mistakenly inscribe as a "register sex offender".*, after the semantic replacement, we can recognize the lemma *"erroneously"* was replaceable by *"mistakenly"*. However, WordNet 2.1 cannot recognize the lemmas *"label"* and *"inscribe"* which can also be replaceable. To address this problem, we simply assumed that two lemmas surrounded by the same context can be replaceable as well. In the experiments, we set the window size of context replacement as 3.

This step is the foundation of the extraction of the sentence different features and can also alleviate the imprecise similarity measure problem existing in STS caused by the possibility of the lemmas in totally different forms sharing the same sense.

3.2 System Configuration

We selected 500 samples from the training data as development set (i.e. test set in SemEval 2012) and performed a series of preliminary experiments to evaluate the effectiveness of different feature types in isolation and also in different combinations. According to the results on the development set, we configured five different systems on each language pair as our final submissions with different feature types and classification strategies. Table1 shows the five configurations of those systems.

System	Feature Set	Description
1	all	flat, SVM
2	best feature sets	flat, SVM
3	best feature sets	flat, Majority Voting
4	best feature sets	flat, only 500 instances for train, SVM
5	best feature sets	hierarchical, SVM

Table 1: System configurations using different strategies based on the results of preliminary experiments.

Among them, System 1 serves as a baseline that used all features and was trained using a flat SVM while System 2 used only the best feature combinations. In our preliminary experiments, different language pairs had different best feature combinations (showed in Table 2). In System 3 we performed a majority voting strategy to combine the results of different algorithm (i.e. MaxEnt, SVM, liblinear) to further improve performance. System 4 is a backup system that used only the training set in SemEval 2012 to explore the influence of the different size of train set. Based on the analysis of the preliminary results on development set, we also find that the misclassification mainly occur between the class of backward and others. So in System 5, we adopted hierarchical classification technique to filter out backward class in the first level using a binary classifier and then conducted multi-class classification among the remaining three classes.

We used a linear SVM with the trade-off parameter C=1000 (also in liblinear). The parameters in SS are set as below: the dimension of sematic space is 100, the weight of missing words is 100 and the regularization factor is 0.01. In the hierarchical classification, we use the liblinear (Fan et al., 2008) to train a binary classifier and SVM for a multi-class classifier with the same parameters in other Systems.

4 Results and discussion

Table 2 lists the final results of our five systems on the test samples in terms of four language pairs. The best feature set combinations for different language pairs are also shown. The last two rows list the results of the best and runner-up team among six participants, which is released by the organizers.

From this table, we have some interesting findings.

Firstly, the feature types BC and SD appear in all best feature combinations. This indicates that the length and sentence difference information are good and effective label indicators.

Secondly, based on the comparison between System 1 and System 2, we find that the behavior of the best feature sets of different language pairs on test and development datasets is quite different. Specifically, the best feature set performs better on FRA-ENG and DEU-ENG data sets than the full feature set. However, the full feature set performs the best on SPA-ENG and ITA-ENG data sets. The reason may be the different distribution properties of test and development data sets.

Thirdly, although the only difference between System 2 and System 4 is the size of training samples, System 4 trained on a small number of training instances even makes a 1.6% improvement in accuracy over System 2 on DEU-ENG data set. This is beyond our expectation and it indicates that the CLTE may not be sensitive to the size of data set.

Fourthly, by adopting a majority voting scheme, System 3 achieves the best results on two data sets among five systems and obtains 45.8% accuracy on FRA-ENG which is the best result among all participants. This indicates the majority voting strategy is a effective way to boost the performance.

Fifthly, System 5 which adopts hierarchical classification technique fails to make further improvement. But it still outperforms the runner-up system in this task on FRA-ENG and DEU-ENG. We speculate that the failure of System 5 may be caused by the errors sensitive to hierarchical structure in hierarchical classification.

In general, our approaches obtained very good results on all the language pairs. On FRA-ENG and DEU-ENG, we achieved the best results among the 16 systems with the accuracy 45.8% and 45.3% respectively and largely outperformed the runner-up. The results on SPA-ENG and ITA-ENG were also promising, achieving the second and third place among the 16 systems.

5 Conclusion

We have proposed several effectively features consisting of sentence semantic similarity and sentence difference, which work together with other features presented by the previous work to solve the cross-lingual textual entailment problem. With the aid of machine translation, we can handle the cross-linguality. We submitted five systems on each language pair and obtained the best result on two data sets, i.e., FRA-ENG and DEU-ENG, and ranked the 2nd and the 3rd on other two language pairs respectively. Interestingly, we find some simple feature types like BC and SD are good class indicators and can be easily acquired. In future work, we will investigate the discriminating power of different feature types in the CLTE task on different languages.

Acknowledgements

The authors would like to thank the organizers and reviewers for this interesting task and their helpful suggestions and comments, which improves the final version of this paper. This research is supported by grants from National Natural Science Foundation of China (No.60903093), Shanghai Pujiang Talent Program (No.09PJ1404500), Doctoral Fund of Ministry of Education of China (No. 20090076120029) and Shanghai Knowledge Service Platform Project (No. ZF1213).

References

Julio Javier Castillo. 2011. A wordnet-based semantic approach to textual entailment and cross-lingual

System	SPA-ENG	ITA-ENG	FRA-ENG	DEU-ENG
1	**0.428**	**0.426**	0.438	0.422
2	0.404	0.420	0.450	0.436
3	0.408	**0.426**	**0.458**	0.432
4	0.422	0.416	0.436	**0.452**
5	0.392	0.402	0.442	0.426
Best feature set	BC+STS+SS +GR+SD	BC+SD+SS +GR+BS	SD+BC+STS	BC+STS+SS +BS+SD
Best	0.434	0.454	0.458	0.452
runner-up	0.428	0.432	0.426	0.414

Table 2: The accuracy results of our systems on different language pairs released by the organizer.

textual entailment. *International Journal of Machine Learning and Cybernetics*, 2(3):177–189.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. In *Proceedings of the PASCAL Workshop on Learning Methods for Text Understanding and Mining*.

Miquel Esplà-Gomis, Felipe Sánchez-Martínez, and Mikel L. Forcada. 2012. Ualacant: Using online machine translation for cross-lingual textual entailment. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 472–476, Montréal, Canada, 7-8 June.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Weiwei Guo and Mona Diab. 2012. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

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 44th Annual Meeting of the Association for Computational Linguistics*, pages 905–912, Sydney, Australia, July.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2012. Soft cardinality+ ml: Learning adaptive similarity functions for cross-lingual textual entailment. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

Elena Lloret, Oscar Ferrández, Rafael Munoz, and Manuel Palomar. 2008. A text summarization approach under the influence of textual entailment. In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008)*, pages 22–31.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Prodromos Malakasiotis and Ion Androutsopoulos. 2007. Learning textual entailment using svms and string similarity measures. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 42–47.

Prodromos Malakasiotis. 2009. Paraphrase recognition using machine learning to combine similarity measures. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pages 27–35.

Yashar Mehdad, Alessandro Moschitti, and Fabio Massimo Zanzotto. 2010a. Syntactic/semantic structures for textual entailment recognition. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1020–1028.

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

Fandong Meng, Hao Xiong, and Qun Liu. 2012. Ict: A translation based method for cross-lingual textual entailment. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 715–720, Montréal, Canada, 7-8 June.

M. Negri, A. Marchetti, Y. Mehdad, L. Bentivogli, and D. Giampiccolo. 2013. Semeval-2013 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

William E Winkler et al. 1999. The state of record linkage and current research problems.