CROSS-LINGUAL SIMILARITY DISCRIMINATION WITH TRANSLATION CHARACTERISTICS

Ying Qin

Department of Computer Science, Beijing Foreign Studies University,
Beijing, China

ABSTRACT

In cross-lingual plagiarism detection, the similarity between sentences is the basis of judgment. This paper proposes a discriminative model trained on bilingual corpus to divide a set of sentences in target language into two classes according their similarities to a given sentence in source language. Positive outputs of the discriminative model are then ranked according to the similarity probabilities. The translation candidates of the given sentence are finally selected from the top-n positive results. One of the problems in model building is the extremely imbalanced training data, in which positive samples are the translations of the target sentences, while negative samples or the non-translations are numerous or unknown. We train models on four kinds of sampling sets with same translation characteristics and compare their performances. Experiments on the open dataset of 1500 pairs of English Chinese sentences are evaluated by three metrics with satisfying performances, much higher than the baseline system.

KEYWORDS

Cross-lingual, Sentence Similarity, Discriminative Model, Imbalanced Training

1. INTRODUCTION

Plagiarism is contrary to the academic integrity. Currently plagiarism detection research draws more attention with increasing plagiarism activities. Translation plagiarism or cross-lingual plagiarism is among the most obfuscated tasks since it can't be found literally. According to the recent report of International Competitions on Plagiarism [1], the precision and recall of cross-lingual plagiarism detection are 0.1 and 0.13 respectively on manually translated data.

Generally plagiarism detection consists of two stages, suspect document retrieval and translation position marking. What we want to focus in the paper is the second stage, that is, given suspect documents in different languages, the system could mark the correspondent sentences of translation.

To find similar sentences in two different languages, most of systems in International Competition on Plagiarism [1-3] resorted to machine translation and compared the translations of machine with the original texts. Reference [4] ever built a statistical word alignment model to detect the similar sentences between English and Spanish.

Although bilingual alignments have achieved good performances [5-6], we don't think it is competent for cross-lingual plagiarism detection. For the translation detection task, it is required to search for the correspondences for a given sentence in an open pool of texts, not a prior aligned sentence set.

DOI : 10.5121/ijaia.2013.4503

To search its translation correspondences for a given sentence, reference [7] and [8] calculated the similarity value of sentence pairs based on semantic distance. The method is named as translation recognition. However multi-lingual thesaurus is always scarce.
In this paper, we'll build a discriminative language model based on Maximum Entropy, combining several translation characteristics to determine cross-lingual similarity between two sentences. The procedure of translation detection is as follows. Firstly the model identifies whether a sentence in source language is similar to the target sentence. Then we rank the positive outputs according to their probabilities. Finally a threshold maybe set to filter out the results with low confidence of judgment, because there is no prior assumption on the existence of translation correspondences in plagiarism detection.

As to the performance of the system, three evaluation metrics are proposed to meet the need of detection in an open set. Two of them are based on accuracy of discrimination, the other is precision of the top-n results after ranking all outputs. We also build a baseline system to compare with.

In the following, Maximum Entropy (ME) model is briefly introduced first. Then we'll depict how to build a similarity discriminative model on simple translation characteristics. Due to the non-translations of a sentence are numerous, question of training ME model on imbalanced data set is discussed in section 4. Experiments and evaluations carried on 1500 pairs of English Chinese sentences are arranged in section 5. As usual, the last section is some conclusions.

2. MAXIMUM ENTROPY MODEL

Maximum Entropy (ME) as a discriminative model making no assumption of unknown facts has been widely applied to nature language processing [9]. It effectively combines different facts to make probability prediction $p(y \mid x)$, in which y is a member of a finite category set γ and x is a member of finite feature information set χ of the target. We here jump over the description of ME for the limit of paper length.

We view the translation correspondence identification as a classification problem. And we simply define the binary similarity between sentences because there are many disputes and divergences about the degree of similarity. Given a sentence, the ME model will discriminate the sentences in other language according to a finite translation feature set χ of them. And prediction y indicates whether the testing sentence is the translation of the given sentence or not, that is, y has a binary value.

3. TRANSLATION CHARACTERISTICS IN ME MODEL

3.1 Universal Translation Characteristics

Linguists had found universal features among translation texts by exploring parallel corpus [10-11], for instance, sentence length ratio of translation pairs is close to the average statistics from large bilingual corpus, divergence of word POS in translation pairs and so on. We are inspired by these findings and try to integrate them into a statistic model to search for the translation pairs.

Length ratio of sentence pairs is one of coarse criteria to predicate translations since long sentences are usually translated into long sentences in another language. According to the bilingual corpus for training, the ratio can be calculated and used as a feature variable f_1 in ME model. Sentence length far from the average number is not likely to be the translation.

Negative sentence is always translated into target language using negations. Sometimes there are several negations in the source sentence. Accordingly there may be more than one negation in the translation. Double negative means positive. Therefore we count the number of negations and use a binary feature to indicate the polarity of the sentence. Generally the polarity of translation pairs should be
same. For example, if there are double negation words in the source sentence, its translation generally contains two or zero negations. If not, it is not likely the translation. The negations or negative expressions in a language are limited to some extent in many languages. In our studies, we collect common negations in English and Chinese to check the polarity of a sentence. In ME model, polarity feature f_2 is also binary, making the model select the candidate sentence with same polarity.

Universal translation characteristics can be used to make prediction from the whole. If we want to determine the similarity of two sentences in different languages, detailed features on the level of words are also required.

3.2 Alignment Features

Shallow statistical features are far from sufficient to make a judgment whether two sentences in different languages are similar or not. Hence we utilize a bilingual lexicon to determine matching meanings on word level. The more matching meanings according to a bilingual dictionary in sentence pairs, the more likely they are translations. To be noticed that functional words carry less meanings comparing to notional words. We divide the words according to their POS and pay more attention on the matching of notional words such as noun, verb and adjective. Features of meaning matching utilized in ME model are real numbers as defined below,

The percentage of matching words including notional and functional words, defined as feature f_3,

$$f_3 = \frac{count\ of\ matching\ words}{count\ of\ words\ in\ sentence}$$

The percentage of matching notional words among all words in the sentence, defined as f_4,

$$f_4 = \frac{count\ of\ matching\ notional\ words}{count\ of\ matching\ words}$$

The percentage of matching notional words to all notional words in the sentence, listed as f_5,

$$f_5 = \frac{count\ of\ matching\ notional\ words}{count\ of\ notional\ words\ in\ sentence}$$

These three features are used in ME model to check the meaning similarity between two sentences according to a bilingual lexicon.

3.3 Bilingual Lexicon Arrangement

Bilingual lexicon is always utilized in cross-lingual studies to compare the meaning similarity on the level of words. However it only provides static meanings without considering the word's contexts in a sentence. So polysemy and sense insufficiency are among the most intractable problems in similarity determination according to a dictionary [12].

To explore word similarity features mentioned in 3.2, we firstly merge several bilingual lexicons to increase the coverage of the lexicon and provide as many translations as possible. However senses in the merged lexicon are listed randomly which will cost more in search and sense disam-

biguation. We try to make minimum error in meaning selection by sorting the translation senses of polysemous words in the lexicon. If the most popular sense is listed front, the error of matching will be lower than doing on random sense order and the efficient will improve at the same time since we always stop matching when coming to the first same.

Some translations of a word in the lexicon are synonyms while some senses are very different. Taking the word *Juvenile* as an example, in the merged English Chinese dictionary, there are totally 13 translation senses in Chinese,

幼稚,少年,少年的,青少年,青少年的,不成熟的,少年读物,适于少年的,年轻的,年幼的,年少的,少女的,适合少年少女的

Generally each Chinese character or Hanzi has meanings and can be viewed as a semantic unit. In the above, the most frequent character in the translations is 少 (which means young) excluding the stop words 的. In fact the primary meaning of Juvenile is young. And other meanings like 幼稚(childish) and 不成熟 (naive) are extensions. Therefore we assume that the more frequent a character is involved in the translations, the more probable it takes the primary meanings of the source word. Based on the assumption, we propose a method to rearrange the translations of the polysemous words in the bilingual lexicon.

For an English polysemy with m Chinese translation senses, each sense X_i is made of n_i Chinese character $C_1...C_{n_i}$, the score of each sense is calculated according to the following formula.

$$S_{X_i} = \sum_{j=1}^{n_i} \frac{f_{C_j}}{mn_i}$$

In which, f_{C_j} is the frequency of character C_j in all senses of this word. Obviously $\sum_{i=1}^{m} S_{X_i} = 1$. The higher score means the more primary meaning the translation owns.

All translations are sorted in descending according to its score. During the mechanical matching, the sense with higher score will be firstly encountered which contains the primary meaning. Then the matching process will stop with success.

After sense arrangement, the order of senses of *Juvenile* is changed into as,
少年,年少的,少年的,青少年的,青少年,年幼的,少女的,年轻的,适于少年的,适合少年少女的,少年读物,不成熟的,幼稚

We can find that the extended meanings are sorted behind.

As to other languages, the semantic unit may take the form of word like English, or other language unit. And the method of sense order arrangement can also be applied to.

4. MODEL TRAINING ON IMBALANCED DATA

Cross-lingual similarity discrimination is viewed as the task of classification in this paper. To train an ME based classifier, standard recipe is collection of samples of all classes firstly. In the translation discriminative model, positive samples are the translations of the given sentence of course. Generally we can acquire the translation pairs from a bilingual corpus as positive samples. However the negative samples or non-translations of a given sentence are unknown and numerous.

To deal with the imbalanced samples in the training of classifiers, three methods are generally used from the data level: undersampling, oversampling and feature selection as reviewed by reference [13]. In document classification, reference [14] built an adaptive SVM classifier learning from one class to handle the extremely imbalanced data. There is no kernel function to transform samples into one class in ME model. So we prepare the training data for ME similarity discriminative model in four ways.

- Cross sampling: Given a set of bilingual sentence pairs, we make cross sampling which only treat the translation of the target sentence as positive sample while all the other sentences in the set as negative samples. For instance, if there are n sentence pairs for training, the number of total samples in training is n^2. Cross sampling is the exhaustive combinations of given training set.

- Reduced sampling: Since the negative samples are much more than the positive in translation discrimination, we then randomly select several negative samples to form training data for reducing the scale of negative samples.

- Undersampling using KNN: Random selection of negative samples can't make the classifier learn from the most difficult samples. Therefore another undersampling method we used in training is to select the most similar negative samples to the positive sample to train the classifier. KNN is applied to select the nearest k negative samples. Distance between the negative and the positive is defined according to the feature values mentioned above.

$$Dis = \sqrt{\sum_i (fn_i - fp_i)^2}$$

In which, fn_i and fp_i denotes the i th feature value of the negative and the positive respectively. Undersampling using KNN make the classifier learn from the hardest samples.

- Synthetic sampling: No one can image the scale and the style of non-translations for a given sentence. Even the cross sampling can't cover the whole negative samples. Therefore we try to make a negative sample which can generalize the features of all non-translations. For each features mentioned above, we average the values of negative samples in cross sampling set to synthesize a negative sample. Then the ratio of positive and negative sample in training is 1:1.

We test the four sampling methods in ME classifiers and find the synthetic sampling runs the best performance on positive samples.

5. EXPERIMENTS

5.1. Data

To build English-Chinese translation discrimination model, we utilize bilingual parallel corpus as experimental data. On Open Source Platform for Chinese NLP [1], 1500 English-Chinese sentence pairs are shared. These sentence pairs are not field-specific. Length of English sentences varies from 3 to 40 words. Hereinafter all the experimental results are 5-fold cross validation, that is, 1200 pairs are used for training the ME classifier and 300 pairs are use for testing.

[1] http://www.nlp.org.cn/docs/download.php?doc_id=991

Bilingual lexicon is merged from 3 English-Chinese dictionaries and Chinese senses of each English word are sorted according the idea mentioned in section 3.2. And the final scale of the lexicon is 144,102 English words.

Three common negation words (不 没 否) are used to detect the polarity of a Chinese sentence. While for an English sentence, we use 28 negation words to check its polarity.
Notional words of English in our experiments refer to noun, verb and adjective according to their POS tagged.

Totally we extract 5 features to train the ME similarity discriminative model.
- Sentence length ratio of English Chinese pairs;
- Percentage of words matching according to the bilingual lexicon;
- Percentage of notional words matching to all words in a sentence;
- Percentage of notional words matching to notional words in a sentence;
- Polarity consistence of sentence pair.

All the features are real value.

5.2. Evaluation metrics

To access the performance of translation discrimination, accuracy is always used which is defined as,

$$ACC_{-all} = \frac{count\ of\ correct\ output}{count\ of\ all}$$

Because all the sentences except its translation are treated as negative samples, for 300 sentence pairs for testing, testing set contains 90000 sentences. Since most of them are non-translation relation, the accuracy may be very high. However what we concern is how many translations if existed are recognized. So we should pay more attention on the performance on real translation pairs or positive samples. Hence we propose another metric of performance ACC-true which is defined as below,

$$ACC_{-true} = \frac{count\ of\ correct\ output\ on\ positives}{count\ of\ all\ positives}$$

In information retrieval, a popular metric used to evaluate IR system is $top\text{-}n$ precision, that is, how many relative results are listed among the first n outputs. Obviously the outputs are ranked according to the probabilities of relevance. Similarly we use the $top\text{-}n$ precision to check whether the real translation of a given sentence is among the first n outputs after ranking. In our experiments the value of n is set 1, 3 and 5. That is three precision metrics are used to evaluate the final performance of the similarity discrimination system, top-1, top-3 and top-5.

5.3. Experimental Results

5.3.1 Sampling of training

Firstly we'd like to show the results of different sampling when training the ME classifiers.

From Table 1 we can see the performance of four sampling methods. In cross sampling, for each sentence there is only one translation, all the other sentences are viewed as negative samples.

Though the ACC_{-all} is rather high, only a small part of real translations (low ACC_{-true}) are picked out rightly. If we randomly select three negative samples from all non-translations and one positive to form training data set, the classifier raise the performance on real translations which the ACC_{-true} reaches 73.67%. Unfortunately if we use the negative samples which are most k nearest to the positive to train ME model (k =1, 3), the performances decrease. The synthetic sampling runs the best among all sampling methods. Even though the ACC-all is little lower than cross sampling, 93.80% of translations of testing sentences are correctly classified. The result is exciting.

Table 1. Performance of Different Sampling in ME Training

Sampling	***ACC-all(%)***	***ACC-true(%)***
Cross	99.67	17.93
Reduced	98.15	73.67
KNN-1	95.65	71.67
KNN-3	98.83	52.00
Synthetic	85.98	93.80

5.3.2 Baseline System

We create a baseline system to compare with ME system. In the baseline system, a given sentence is firstly translated by looking up to the same English-Chinese dictionary as ME system, then simply using word-matching technique to find its translations. According to the degree of similarity, we sort the target sentences and select the positive translations.

Sentence similarity based on mechanic word-matching is calculated as the follows,

$$SenSim = \frac{count\ of\ words\ matched}{count\ of\ words\ in\ two\ sentences}$$

5.3.3 Final performance

In open testing, existence of translation for a given sentence is not assumed. Therefore we rank all the sentences which are discriminated as translation similarities according to the probabilities to check if the real translation is listed among the $top\text{-}n$ results. Figure 1 shows the top-1, top-3 and top-5 precision of the final outputs on 300 sentence pairs. Among the ranked outputs, the top-1 precision reaches 76.13%. And 86.33% translation of a testing sentence is listed among the top-5 outputs. The performances of the baseline system are also shown in Figure 1. Obviously ME system outperforms the baseline.

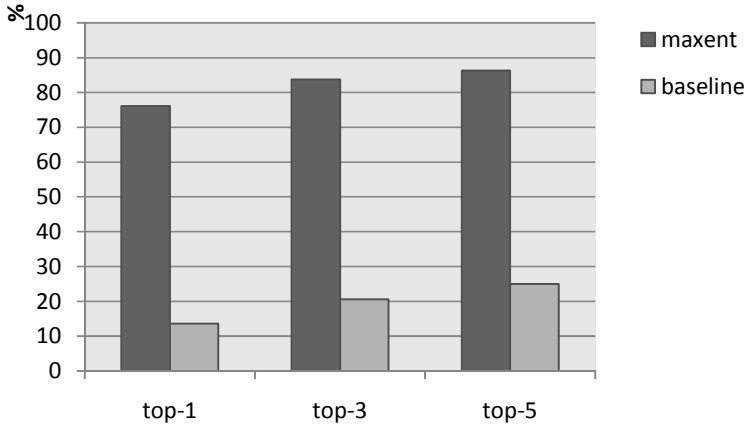

Figure 1. Top-n precision of ME comparing with the baseline

In applications, we can further use threshold to limit the outputs of discrimination in open cross-lingual translation detection when if there is no translation of the given sentence in testing set, no suspect sentence will be output. On the observation of experimental bilingual corpus, the threshold is set as 0.522 empirically.

5.3.4 Error analysis

We analyze the sentence pairs which are detected incorrectly. There are three kinds of causes. Firstly they have similar general translation features like similar length, same polarity. Another reason is due to the coverage of the bilingual lexicon. It is crucial to the value of feature f_3 - f_5. If a word can't be translated according to the lexicon, it is same to a mismatched word. And this doesn't hold. Also the features used in ME classifier are not rich. We'll further explore more translation features.

6. CONCLUSIONS

To recognize translation sentence pairs, we build a classifier based on Maximum Entropy to discriminate sentences in a different language using simple translation characteristics such as the percentage of matching words according to a bilingual lexicon, sentence length ratio, polarity consistence of sentence pair and so on. Accuracy of the classifier on positive samples can reach 93.80% when using synthetic training samples. And the $top-n$ precision of ranked outputs on all testing data is also satisfying comparing to the baseline. In cross-lingual plagiarism detection, the similarity discrimination model could be used to mark the translation correspondences for each target sentence.

Translation characteristics are universal. In this experiment we only explore 5 features. Next we'll integrate more translation features into ME model. Also we'll test the translation discrimination system on more data.

ACKNOWLEDGMENTS

The work is supported by the BFSU Special Fund for Scientific Research (No. 2009JJ056). Thanks are due to the support of National Research Centre for Foreign Language Education, BFSU. Acknowledgments are also given to anonymous reviewers who gave good advices to the paper writing.

REFERENCES

[1] Martin Potthast, Andreas Eiselt, Benno Stein, and Paolo Rosso (2010). Overview of the 2nd International Competition on Plagiarism Detection. In Martin Braschler and Donna Harman, editors, Notebook Papers of CLEF 10 Labs and Workshops.

[2] Martin Potthast,et al. (2009). Overview of the 1st International Competition on Plagiarism Detection. In Benno Stein, Paolo Rosso, Efstathios Stamatatos, Moshe Koppel, and Eneko Agirre, editors, SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), p1-9.

[3] Martin Potthast, et al. (2011). Overview of the 3rd International Competition on Plagiarism Detection. Notebook Papers of CLEF 2011 Labs and Workshops, p19-22.

[4] Alberto Barrón-Cedeño, Paolo Rosso，Eneko Agirre , Gorka Labaka. (2010). Plagiarism Detec-tion across Distant Language Pairs. Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 37–45.

[5] Peter F Brown, Jennifer C Lai, Robert L Mercer (1991). Aligning sentences in parallel corpora. In Proceedings of the 47 Annual Meeting of the ACL.

[6] Franz Josef Och, Hermann Ney (2003). A Systematic Comparison of Various Statistical Align-ment Models. Computational Linguistics, vol. 29, pp. 19-51.

[7] Steinberger R., Pouliquen B., Hagman J. (2002). Cross-lingual Document Similariy Calculation Using the Multilingual Thesaurus EUROVOC. A. Gelbukh (Ed.): CICLing 2002, LNCS 2276, pp. 415-424.

[8] Smith Noah A.(2002). From Words to Corpora: Recognizing Translation. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 95-102.

[9] Berger A L. (1996). A Maximum Entropy Approach to Natural Language Processing. Computational Linguistic, 22 (1): pp. 39-71.

[10] Hu X.Y and Zeng J.(2011). New Trends in the Corpus-based Research on Translation Univer-sals. Journa l of PLA University of Foreign Languages. 2011(1):56-62.

[11] Mauranen, A. (2008). Universal tendencies in translation. Gunilla Anderman, Margaret Rogers. Incorporating Corpora: The Linguist and the Translator. Clevedon: Multilingual Matters Chen J.P. 2006. A Lexical Knowledge Base Approach for English–Chinese Cross-Language Informa-tion Retrieval. Journal of the American Society for Information Science and Technology, 57(2):233–243.

[12] Chen J.P. (2006). A Lexical Knowledge Base Approach for English–Chinese Cross-Language Information Retrieval. Journal of the American Society for Information Science and Technol-ogy, 57(2):233–243.

[13] Kotsiantis Sotiris, Kanellopoulos Dimitris, Pintelas Panayiotis. (2006). Handling imbalanced datasets: A review. International Transactions on Computer Science and Engineering, 30(1):25-36.

[14] Manevitz Larry M. and Yousef Malik. (2001). One-Class SVMs for Document Classification. Journal of Machine Learning Research 2 .pp.139-154.

Author

Ying Qin received her doctor degree from Beijing University of Posts and Telecommunications in 2008. Her major research interest is natural language processing and computational linguistics. Her Google Scholar Profile link is http://scholar.google.com/citations?hl=en&user=Cum7cAkAAAAJ&view_op=list_works&gmla=AJsNF4EivnaStOqjjI0l3Cu9yzzs5lWlEpty1220fpqdV05CzqhlLPmp8C8t_mieh6XChu3Th6ufCCl5oS-XM5s6Q4DF6DZH2_J33EQqCtWq14NVj-7oYFl7w-HLa09FQXj-TTQY5q