Multi-feature Based Chinese-English Named Entity Extraction from Comparable Corpora

Min Lu, Jun Zhao

[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100080, China
{ mlu, jzhao} @nlpr.ia.ac.cn

Abstract. Bilingual Named Entity Extraction is important to some cross language information processes such as machine translation (MT), cross-lingual information retrieval (CLIR), etc. A lot of previous work extracted bilingual Named Entities from parallel corpus. Here we propose a multi-feature based method to extract bilingual Named Entities from comparable corpus. We first recognize the Chinese and English Named Entities respectively from the Chinese and English part of the comparable corpus. Then all the feature scores are calculated for every possible pair of Chinese and English Named Entities. At last we combine these feature scores together and decide which pairs are mutual translations. For translation score calculation, we didn't use the formula of IBM model 1 like previous approach. In stead, we used a modified edit distance to take the order of words into consideration. Experiment shows that the F-score of this method increased by 11 %. And with the multi-feature integration strategy encouraging results are obtained.

1 Introduction

Named Entities (NE) such as person names, location names, organization names, etc, carry essential information in human language [1]. Examples of some base Named Entities are as follows:

Person name: 胡锦涛 Hu jintao 迈克尔·欧文 Michael Owen

Location name: 渥太华 Ottawa 华盛顿 Washington

Organization name: 教育部 Ministry of Education 国家安全委员会 National Security Council

Bilingual Named Entity Extraction is very important to many cross language information processes such as machine translation (MT), cross-lingual information retrieval (CLIR), etc. In machine translation systems we translate the NEs in the sentences according to a bilingual NE list. Thus the list should be constantly updated to ensure correct translation of new NEs, since new NEs appear on the Internet everyday. This paper is about how to extract bilingual NEs from Comparable Corpora, which can be easily obtained from the Internet.

A lot of work has been done to get bilingual NEs from parallel corpora, aligning English NEs and Chinese NEs in parallel sentences [2][3]. Although encouraging results have been obtained, the parallel corpus required in this approach is not easy to obtain. The construction of the parallel corpus costs huge human work and a long time.

Compared with parallel corpus, comparable corpus can be obtained much more easily. Comparable corpus is one that contains non-sentence-aligned, non-translated bilingual documents that are topic-aligned. For example, newspaper articles from two sources in different languages, within the same window of published dates, can constitute a comparable corpus [4].

However, the bilingual NE extraction from comparable corpus is much more difficult than that from the parallel corpus. In parallel corpus, given a source NE to be translated in a source sentence, the translation candidates are only the NEs in the corresponding target sentence. But in comparable corpus the possible translations for a Chinese NE are NEs of the same type in the whole target language corpus. The search space is much larger than that in the parallel corpus, which largely increases the difficulty of

extraction. In our work, we propose a multi-feature integration strategy to try to solve this problem. Multi-features including phonetic feature, semantic feature, length feature and context feature can gather more similarity in various aspects to find the correct NE pairs from large amount of candidates. Experimental results show that all these features are very useful for the task.

The rest of this paper is organized as follows: In section 2, we discuss related work on bilingual NE extraction. Section 3 introduces the approach we use, including the features we selected. Experimental results and analysis are shown in section 4. At last, conclusion is given in Section 5.

2. Related Work

Most previous methods of extracting bilingual NEs are based on parallel corpora. [2][3] integrated multi-features including transliteration feature, translation feature and other features in parallel corpus to align Named Entities in parallel sentences. In the process, NEs are first recognized from the source language and target language respectively. Then, NEs in each pair of parallel sentences are aligned according to their features. [5] extracted formulation and transformation rules of multilingual named entities from multilingual named entity corpora and used them to CLIR. However, the corpus they used is not easy to obtain, while comparable corpus we used is much more accessible.

[6] adopted the context co-occurrence relations to extract bilingual lexicon in non-parallel corpora which was one of some early researches on non-parallel corpora. This approach is based on the assumption that if two words in different languages have similar contexts, it is most likely that they are mutual translations. However, it was not very effective with an accuracy of 30% when only the top one candidate was considered. [7] integrated this method and transliteration to extract keywords in comparable corpus. This task is similar to ours. However, it only considered two features while ours combine more.

3. Named Entity Extraction

3.1 Feature Functions Selection

3.1.1 Transliteration Score

Person names and most location names in source language are similar in their pronunciations with those in the target language. Transliteration feature is used to characterize this property. Transliteration is the process of replacing words in the source language with their approximate phonetic or spelling equivalents in the target language [8]. Most previous approaches resorted to phoneme similarity, where a pronunciation lexicon is needed. Fei Huang [3] constructed a transliteration model on the surface level which didn't need the pronunciation. This method uses pinyin as intermediate, which is the Romanized representation of Chinese characters. It had two levels of transliteration, Chinese character to pinyin syllable and pinyin syllable to English letter string. Our transliteration probability from English letters to pinyin letters is trained based on this thought. We view pinyin as the source language and English as the target language, and train the probability from English letters to pinyin letters using the IBM Model 4 on a LDC bilingual person name list.

To measure the similarity between a pinyin string and an English string, we compute the edit distance between them. The standard cost function in edit distance is 1 or 0. We use transliteration probability as the cost function to acquire the phonetic distance. The cost function is:

$$1-\left(p(r_i \mid e_j)+p(e_j \mid r_i)\right)/2 \qquad (1)$$

where r_i and e_j are the i^{th} pinyin letter and the j^{th} English letter respectively.

We normalize the edit distance according to the string length and convert it to the similarity between pinyin string and English string. So the transliteration score is:

$$Score_{trasli}(ne_c, ne_e) = \exp(-\frac{d}{L_{larger}}) \quad (2)$$

where d is the edit distance and L_{larger} is the length of longer string.

3.1.2 Translation Score

Some Named Entities contain not only phonetic similarity but also semantic similarity. For example, in the location name "小落矶山脉" and "Little Rocky Mountains", "落矶" and "Rocky" are phonetically similar, while "小" and "山脉" are the semantic translations of "Little" and "Mountains" respectively. For most Chinese organization names and their corresponding English ones, the words composing them are mutual translations, which is shown in table 1. Thus it is needed to consider the translation probability between words in English NE and words in Chinese NE to reckon in the semantic similarity.

Table 1. Example of organization names

"欧洲委员会"	" European Commission"
"海关总署"	"General Administration of Customs"
"教育部"	"Education Ministry"
"国家安全委员会"	"National Security Council"

Previous work extracting bilingual NEs from parallel corpus defined the translation score using the IBM model 1 or a similar formula. They didn't consider the order of words in NEs. In fact the regulation of word order for NEs is relatively simple. Consider the organization names. Some Chinese organization names have the same order with their English names, while others have the opposite order when there is a word "of" in the middle of the English organization names. So we can use a modified edit distance to evaluate the similarity between Chinese NEs and English NEs. When the English NE has a word "of" on the center, we swap the words on the left of "of" with those on the right. Our cost function is $1 - p(e_j \mid c_l)$, where the translation probability $p(e_j \mid c_l)$ can be estimated from a large parallel corpus using IBM-mode l [9]. Assume that the edit distance between two words is d and the larger word length is L_{larger}, then our translation score is given below:

$$Score_{trans}(ne_c, ne_e) = \exp(-d / L_{larger}) \quad (3)$$

Experiment demonstrates that our translation score considering the word order performs better than previous methods.

3.1.3 Word Length Score

Word length score represents the length relationship between the source NE and the target NE. Statistics on a bilingual organization name list with 25,380 pairs shows that 74% of organization name pairs have length measure more than 0.7. Here length measure refers to the ratio of the shorter NE's length to its translation's length. For the NEs which need transliteration like person names, we first transform them into pinyin, and then compare the pinyin strings with their corresponding English NEs. Statistics on a person name list (672,638 pairs) indicates that pairs with length measure larger than 0.7 occupy 85.6% in all person name pairs. So it is reasonable to assume that a NE and its translation should be comparative in length, except for the abbreviation. The word length score is defined as:

$$Score_{length}(ne_c, ne_e) = \frac{L_{shorter}}{L_{larger}} \quad (4)$$

where $L_{shorter}$ is the length of the shorter NE in a pair, and L_{larger} is the length of the longer one. For organization names, the length refers to the number of the words composing them. And for location names and person names the length is the number of characters in the English NE or pinyin string converted from the Chinese NE. The more comparative the length of the two NEs in different language is, the higher the score is.

3.1.4 Context Score

Pascale Fung [6] used context information to extract bilingual lexicon. It assumes that the words in the source and target language are likely to be mutual translations if their context is similar. Based on the assumption, the standard approach builds a context vector respectively for the source and target word. Then the context vector of the source word is translated to the target language, so that we can compare the source context vector with the target context vector and a similarity between them is also calculated. The detail algorithm is as follows:

a. For a Chinese NE ne_c, we collect its context words in the whole Chinese corpus, that is, the words in a specified window around ne_c. Then we construct a Chinese context vector vec_c for ne_c and put all the context words included in the bilingual lexicon in it (We'll map the Chinese context vector to English context vector using this lexicon). The dimension of the vector is the same with the number of entries the lexicon has. And each term (represented as t_i) in the context vector vec_c corresponds to one entry in the lexicon. The value of the term (represented as w_i) denotes the term's importance to the Chinese NE ne_c. The weight w_i can be calculated as:

$$w_i = tf_i * \log\left(\frac{n+1}{n_i}\right) \quad (5)$$

where tf_i is the frequency that t_i appears as a context word of ne_c, n_i is the number of different Chinese NEs around which t_i has appeared, and n is the total number of different Chinese NEs. Accordingly, if the word corresponding to a certain term doesn't appear around ne_c, the value w_i is zero.

b. For an English NE ne_e, we can acquire a context vector vec_e in a similar manner as the Chinese NE.

c. Map the Chinese context vector vec_c to a English vector vec_{trans} with a bilingual lexicon by translating the key of each term in Chinese context vector to English. The lexicon we use is the LDC Chinese-English lexicon with 54170 entries.

d. Compare the two context vector vec_{trans} and vec_e, and calculate the similarity between them.

Based on the above algorithm, the context score is obtained as:

$$Score_{\cos}(ne_c, ne_e) = \frac{vec_{tran} \cdot vec_e}{|vec_{tran}||vec_e|} \quad (6)$$

3.2 Bilingual NE Extraction

3.2.1 Named Entities Recognition

The process of bilingual NE extraction is like this: first we recognize the Chinese Named Entities in the Chinese corpus with the Chinese NE tagger NLPRCSegTagNer, and recognize the English Named Entities in the English part using GATE. A NE pair can be composed by any two NEs coming respectively from the Chinese NEs and English NEs we've recognized. Then for each NE pair we decide whether they are mutual translations according to their feature scores.

3.2.2 Multi-Feature Integration

A weighted sum of the scores is given below by combining all the features mentioned above.

$$Score_{all}(ne_c, ne_e) = \mu_1 Score_{traslit}(ne_c, ne_e) + \mu_2 Score_{trans}(ne_c, ne_e) + \mu_3 Score_{length}(ne_c, ne_e) + \mu_4 Score_{context}(ne_c, ne_e) \quad (7)$$

where μ_1, μ_2, μ_3, μ_4 are the weights for each feature, which are empirically chosen based on experiment.

4. Experimental Results and Discussion

4.1 Corpus, Test Set and Evaluation Method

Corpus

We take Chinese and English news stories in the same period downloaded from the Internet as our corpus. The Chinese part of the corpus contains the news published in 2005 from the Chinese version of People's Daily and Sina network. The English part includes news report in the same year from the English version of People's Daily and Chinadaily. Since the Chinese part and the English part of the corpus share many identical topics, and they are not mutual translations, this corpus belongs to the comparable corpus.

The size of the Chinese corpus is about 469M with about 63,000,000 words, and the size of the English corpus is about 264M with about 28,000,000 words.

Test Set

We first segmented the Chinese text and tokenize the English text, and then tagged them respectively. After that, NEs were recognized respectively from the Chinese and English part of the corpus. We obtained 98,107 Chinese Person names, 55,591 English Person names, 26,167 Chinese Location names, 51,166 English Location names, 63,010 Chinese Organization names and 13,300 Organization names.

So many NEs were acquired that the evaluation became difficult since we can not manually count how many NE pairs are mutual translations. In order to reduce the size of test data and facilitate our test, we selected the Chinese NEs which occurred at least 10 times and had a translation with frequency larger than 10 in the English corpus as our test set. We call these words Chinese source words. Then we selected the English NEs with frequency above 10 in the English corpus as the English candidate words. The answer set is the NE pairs we can find in the LDC Named Entity Dictionary.

Evaluation Method

As person names, location names and organization names have different characters, they are processed respectively with different weights.

We first calculate all feature scores of each possible NE pair in the test set, and then using those scores we calculate a total score using the formula (7). After that, M NE pairs with the highest scores were selected out. At last, for every Chinese NE, N NE pairs are chosen from the M pairs as results of our system, that is, these NE pairs are regarded as mutual translations. Assume that our system selects R_1 pairs of NE as the final results, among which R_2 ($R_2 \le R_1$) is real mutual translations, and the test set actually contains R_3 pairs of NE. So our precision is $P = \frac{R_2}{R_1}$. The recall is $R = \frac{R_2}{R_3}$. And the F-score is $F = \frac{2 \times P \times R}{P + R}$.

4.2 Experiment

Three sets of experiments are carried out to investigate the performance of the multi-feature integration method. Firstly we compare our proposed method with previous method of calculating translation scores. Secondly, we test how each feature influences the system's performance. At last, the effect of varying the parameter M is investigated.

4.2.1 Comparing Different Methods Adopted in Translation Score Calculation

Previous work used IBM model 1 or similar formula to calculate the translation score. These methods didn't take order of words into account. We use a modified edit distance to consider both the semantic similarity and the word order and acquired a much better result.

In table 2 our method and method in [2] are compared on the organization name test set. The translation score in [2] is defined as:

$$S_{trans}(ne_c, ne_e) = \sum_{j}^{m} \sum_{i}^{n} p(c_j \mid e_i) \quad (8)$$

where c_j, e_i is the j^{th} Chinese word and the i^{th} English word respectively, m and n is the length of Chinese NE and English NE respectively.

Table 2. Comparing edit distance with previous method (The parameter is M = 10000, N = 1)

system	P	R	F
previous	33.1%	25.6%	28.9%
edit distance	**45.0%**	**36.2%**	**40.1%**

The result indicates that modified edit distance method greatly outperforms Donghui Feng's method [2] on translation score calculation. It leads to about 11% increase in F-score.

We also tried Fei Huang's approach [3], which use formula of IBM model 1 as the translation score. However it worked worse since multiplication in it and sparse data made almost the whole score is zero.

4.2.2 Influence of Each Feature

In order to investigate the influence by each feature, we add them one by one to the system, and view the change of performance. Table 3 shows the precision/recall/F-score using different feature sets, where the parameter is M = 10000, N = 1.

Table 3. Performance of the system when different feature sets are used

features selected	P	R	F
Person Name			
transliteration	53.0%	46.7%	49.6%
transliteration+context	60.3%	52.8%	56.3%
transliteration+context+length	61.7%	54.7%	58.0%
transliteration+context+length+translation	**69.9%**	**62.2%**	**65.8%**
Location Name			
transliteration	44.4%	33.0%	37.9%
transliteration+context	74.4%	55.5%	63.6%
transliteration+context+length	75.8%	57.0%	65.0%
transliteration+context+length+translation	**80.0%**	**60.9%**	**69.2%**
Organization Name			
transliteration	3.1%	2.5%	2.8%
transliteration+context	36.5%	29.2%	32.4%
transliteration+context+length	37.1%	29.7%	33.0%
transliteration+context+length+translation	**66.3%**	**53.3%**	**59.1%**

It can be seen that by adding more information, both precision and recall are improved. So every feature is useful. Especially transliteration score and context score is more effective than other feature scores for person name and location name, while translation score and context score lead to more improvement for organization name.

4.2.3 Influence of Parameters

We also investigate the effect of varying M. The results are shown in Table 4. One can see that when M increases, the precision becomes lower while the recall becomes higher.

Table 4. Effect when M is changed

M	P	R	F	P	R	F	P	R	F
	Person Name			Location Name			Organization Name		
1000	85.0%	50.0%	62.9%	89.6%	48.2%	62.7%	74.2%	46.2%	57.0%
5000	72.2%	60.0%	65.5%	82.1%	59.7%	69.1%	66.2%	52.3%	58.4%
10000	69.9%	62.2%	65.8%	80.0%	60.9%	69.2%	66.3%	53.3%	59.1%

5. Conclusions

We propose a multi-feature based method to extract bilingual NE pairs from comparable corpus, which is harder than extracting them from parallel corpus like a lot of previous work. When calculating the translation score, a modified edit distance method is used, which is proved more effective than previous method. Experiment on one year's news comparable corpus shows that our multi-feature method gets encouraging results.

Acknowledgments. This work was supported by the Natural Sciences Foundation of China under grant No. 60372016, the Natural Science Foundation of Beijing under grant No. 4052027. Thanks to all the persons that helped us, especially Wang Gen, Cai Xunliang, Li Xiaocui, Wu Youzheng, Liu Feifan, Duan Xiangyu.

References

1. Hobbs, J. et al. 1996. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural Language Text, MIT Press. Cambridge, MA.
2. Dong-Hui Feng, Ya-Juan Lv, Ming Zhou,"A New Approach for English-Chinese Named Entity Alignment", 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, Jul. 2004.
3. Fei Huang, Stephan Vogel and Alex Waibel, Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-feature Cost Minimization, in the Proceedings of the 2003 Annual Conference of the Association for Computational Linguistics (ACL'03), Workshop on Multilingual and Mixed-language Named Entity Recognition, July, 2003
4. Pascale Fung and Percy Cheung, Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM, In Proceedings of EMNLP 2004, Barcelona, Spain: July 2004.
5. Hsin-Hsi Chen, Changhua Yang and Ying Lin (2003). "Learning Formulation and Transformation Rules for Multilingual Named Entities." Proceedings of ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models, July 12, Sapporo, Japan, 2003, 1-8.
6. Pascale Fung, “Statistical View on Bilingual Lexicon Extraction: from Parallel Corpora to Non-parallel Corpora”. In Lecture Notes in Artificial Intelligence, Springer Publisher, 1998, vol 1529, 1-17. Invited speech, AMTA 98.
7. Li Shao, Hwee Tou Ng (2004). “Mining New Word Translations from Comparable Corpora.” COLING 2004.
8. Y. Al-Onaizan and K. Knight. 2002. Named Entity Translation: Extended Abstract. In Proceedings of Human Language Technology 2002, pp. 111-115, San Diego, CA, March, 2002.
9. P. F. Brown, S. A. Della Pietra, V. J. Della Pietra and R.L. Mercer. The mathematics of Machine Translation: Parameter Estimation. In Computational Linguistics, vol 19, number 2. pp263-311, June, 1993.