English-Chinese Bi-Directional OOV Translation based on Web Mining and Supervised Learning

Yuejie Zhang, Yang Wang and Xiangyang Xue
School of Computer Science
Shanghai Key Laboratory of Intelligent Information Processing
Fudan University, Shanghai 200433, P.R. China
{yjzhang,072021176,xyxue}@fudan.edu.cn

Abstract

In Cross-Language Information Retrieval (CLIR), Out-of-Vocabulary (OOV) detection and translation pair relevance evaluation still remain as key problems. In this paper, an English-Chinese Bi-Directional OOV translation model is presented, which utilizes Web mining as the corpus source to collect translation pairs and combines supervised learning to evaluate their association degree. The experimental results show that the proposed model can successfully filter the most possible translation candidate with the lower computational cost, and improve the OOV translation ranking effect, especially for popular new words.

1 Introduction

In Cross-Language Information Retrieval (CLIR), most of queries are generally composed of short terms, in which there are many Out-of-Vocabulary (OOV) terms like named entities, new words, terminologies and so on. The translation quality of OOVs directly influences the precision of querying relevant multilingual information. Therefore, OOV translation has become a very important and challenging issue in CLIR.

The translation of OOVs can either be acquired from parallel or comparable corpus (Lee, 2006) or mining from Web (Lu, 2004). However, how to evaluate the degree of association between source query term and its target translation is quite important. In this paper, an OOV translation model is established based on the combination pattern of Web mining and translation ranking. Given an OOV, its related information are gotten from search results by search engine, from which the possible translation terms in target language can be extracted and then ranked through supervised learning such as Support Vector Machine (SVM) and Ranking-SVM (Cao, 2006). The basic framework of the translation model is shown in Figure 1.

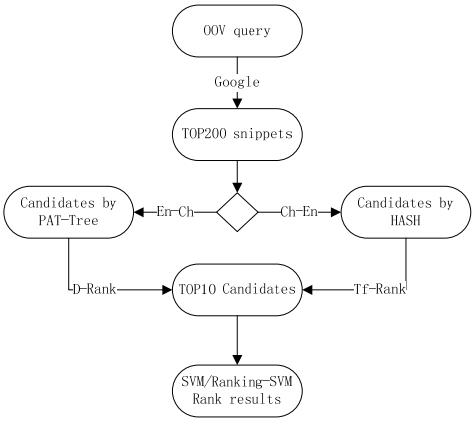

Figure 1. The basic framework of English-Chinese Bi-Directional OOV translation model.

2 Related Research Work

With the rapid growth of Web information, increasing new terms and terminologies cannot be found in bilingual dictionaries. The state-of-art OOV translation strategies tend to use Web itself as a big corpus (Wang, 2004; Zhang, 2004). The quick and direct way of getting required information from Web pages is to use search engines, such as Google, Altavista or Yahoo. Therefore, many OOV translation models based on Web mining are proposed by researchers (Fang, 2006; Wu, 2007).

By introducing supervised learning mechanism, the relevance between original OOV term and extracted candidate translation can be accurately evaluated. Meanwhile, the model proposed exhibits better applicability and can also be applied in processing OOVs with different classes.

3 Chinese OOV Extraction based on PAT-Tree

For a language that has no words boundary like Chinese, PAT-Tree data structure is adopted to extract OOV terms (Chien, 1997). The most outstanding property of this structure is its Semi Infinite String, which can store all the semi-strings of whole corpus in a binary tree. In this tree, branch nodes indicate direction of search

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 129–132,
Suntec, Singapore, 4 August 2009.

and child nodes store information about index and frequency of semi infinite strings. With common strings being extracted, large amounts of noisy terms and fragments are also extracted. For example, when searching for the translation of English abbreviation term "*FDA*", some noisy Chinese terms are extracted, such as "国食品" (17 times), "美国食品" (16 times), "美国食品药" (9 times). In order to filter noisy fragments, the simplified Local-Maxima algorithm is used (Wang, 2004).

4 Translation Ranking based on Supervised Learning

4.1 Ranking by Classification and Ordinal Regression

Based on the extracted terms, the correct translation can be chosen further. A direct option is to rank them by their frequency or length. It works well when the OOV term has a unique meaning and all the Web snippets are about the same topic. However, in much more cases only the highly related fragments of OOV terms can be found, rather than their correct translations. To evaluate the relevance of translation pair precisely, SVM and Ranking-SVM are employed as classifier and ordinal regression model respectively.

4.2 Feature Representation

The same feature set is utilized by SVM and Ranking-SVM.

(1) *Term frequency*: f_q denotes the frequency of OOV to be translated in all the Web snippets of search results. tf_i indicates the number of the translation candidate in all the snippets. df_i represents the number of Web snippets that contains the candidate. df_t means the number of snippets that contains both OOV to be translated and the candidate.

(2) *Term length*: *Len*( ) is the length of the candidate.

(3) *Cooccurrence Distance*: *C-Dist* is the average distance between the OOV query and the translation candidate, computed as follows.

$$C\text{-}Dist = \frac{Sum(Dist)}{df_t} \quad (1)$$

where *Sum*(*Dist*) is the sum of distance in each translation pair of every snippet.

(4) *Length Ratio*: This is the ratio of OOV query length and translation candidate length.

(5) *Rank Value*:

i. *Top Rank* (*T-Rank*): The rank of snippet that first contains the candidate. This value indicates the rank given by search engine.

ii. *Average_Rank* (*A-Rank*): It is the average position of candidate in snippets of search results, shown as follows.

$$A - Rank = \frac{Sum(Rank)}{df_i} \quad (2)$$

where *Sum*(*Rank*) denotes the sum of every single rank value of snippets that contains the candidate.

iii. *Simple_Rank* (*S-Rank*): It is computed based on $Rank(i)=tf_i*Len(i)$, which aims at investigating the impact of these two features on ranking translation.

iv. *R-Rank*: This rank method is utilized as a comparison basis, computed as follows.

$$R - Rank = \alpha \times \frac{|S_n|}{L} + (1-\alpha) \times \frac{f_n}{f_{OOV}} \quad (3)$$

where α is set as 0.25 empirically, $|S_n|$ represents the length of candidate term, L is the largest length of candidate terms, f_n is tf_i, and f_{oov} is f_q in Feature (1).

v. *Df_Rank* (*D-Rank*): It is similar to *S-Rank* and computed based on $Rank(i)= df_i *Len(i)$.

(6) *Mark feature*: Within a certain distance (usually less than 10 characters) between the original OOV and candidate, if there is such a term like "全称", "中文叫", "中文译为", "中文名称", "中文称为", "或称为", "又称为", "英文叫", "英文名为", this feature will be labeled as "+1", else "-1" instead.

Among these features above, some features come from search engine like (1) and (5) and some ones from heuristic rules like (3) and (6). Through the establishment of feature set, the translation candidate can be optimized efficiently and the noisy information can also be filtered.

5 Experiment and Analysis

5.1 Data Set

For the performance evaluation of Chinese-English OOV translation, the corpus of NER task in SIGHAN 2008 provided by Peking University is used. The whole corpus contains 19,866 person names, 22,212 location names and 7,837 organization names, from which 100 person names, 100 location names and 100 organization names are selected for testing. Meanwhile, 300 English named entities are chosen randomly from the terms of 9 categories, which include movie name, book title, organization name, brand name, terminology, idiom, rare animal name, person name

and so on. These new terms are used as the testing data for English-Chinese OOV translation.

5.2 Evaluation Metrics

Three parameters are used for the evaluation of translation and ranking candidates.

$$N-Inclusion-Rate = \frac{number\ of\ correct\ translation\ in\ top\ N\ translations}{total\ number\ of\ OOV\ terms\ to\ be\ translated} \quad (4)$$

$$R-Precision(term_i) = \frac{number\ of\ correct\ transaltion\ in\ top\ R\ translations}{number\ of\ correct\ translations\ for\ term_i\ to\ be\ translated} \quad (5)$$

$$R-Precision = \frac{\sum_{i=1}^{T} R-Precision(term_i)}{total\ number\ of\ OOV\ terms\ to\ be\ translated} \quad (6)$$

where T denotes the number of testing entities. The first one is a measurement for translation and the others are used for ranking measurement.

5.3 Experiment on Parameter Setting

Frequency and length are two crucial features for translation candidates. To get the most related terms into top 10 before the final ranking, a pre-rank testing is performed based on *S-Rank*, *R-Rank* and *D-Rank*. It can be seen from Figure 2 that the pre-rank by *D-Rank* exhibits better performance in translation experiment.

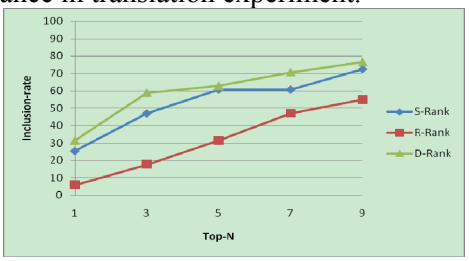

Figure 2. The impact of different *Pre-Rank* manners on English-Chinese OOV translation.

In search results, for some English OOV terms such as "*BYOB*(自带酒水)", there are few candidates with better quality in top 20 snippets. Therefore, in order to find how many snippets are suitable in translation, the experiment on snippet number is performed. It can be observed from Figure 3 that the best performance can be obtained by utilizing 200 snippets.

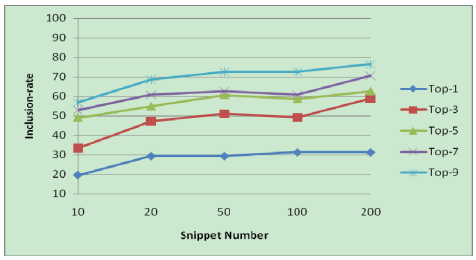

Figure 3. The impact of different snippet number on English-Chinese OOV translation.

5.4 Experiment On English-Chinese Bi-Directional OOV Translation

The experimental results on 300 English new terms are shown in Table 1.

N-Inclusion-Rate	English-Chinese OOV Translation
Top-1	0.313
Top-3	0.587
Top-5	0.627
Top-7	0.707
Top-9	0.763

Table 1. The experimental results on English-Chinese OOV translation.

The experimental results on 300 Chinese named entities are shown in Table 2.

N-Inclusion-Rate	Person Name	Location Name	Organization Name
Top-1	0.210	0.510	0.110
Top-3	0.390	0.800	0.280
Top-5	0.490	0.900	0.400
Top-7	0.530	0.920	0.480
Top-9	0.540	0.930	0.630

Table 2. The experimental results on Chinese-English OOV translation.

It can be observed from Table 2 that the performance of Chinese location name translation is much higher than the other two categories. This is because most of the location names are famous cities or countries. The experimental results above demonstrate that the proposed model can be applicable in all kinds of OOV terms.

5.5 Experiment on Ranking

In SVM-based and Ranking-SVM-based ranking experiment, the statistics on training data are shown in Table 3. For SVM training data, the "*Related*" candidates are neglected. The experimental results on ranking in English-Chinese and Chinese-English OOV translation are shown in Table 4 and 5 respectively.

Number of Candidates	*Correct*	*Related*	*Indifferent*
English-Chinese	234	141	250
Chinese-English	240	144	373

Table 3. Statistics of training data for ranking.

English-Chinese	Top-1 *Inclusion*	Top-3 *Inclusion*	*R-Precision*
D-Rank	0.313	0.587	0.417
T-Rank	0.217	0.430	0.217
SVM	0.530	0.687	0.533
Ranking-SVM	**0.550**	**0.687**	**0.547**

Table 4. The experimental results on ranking in English-Chinese OOV translation.

Chinese-English	Top-1 *Inclusion*	Top-3 *Inclusion*	*R-Precision*
TF-Rank	0.277	0.490	0.287
T-Rank	0.197	0.387	0.207
SVM	0.347	0.587	0.347
Ranking-SVM	**0.357**	**0.613**	**0.387**

Table 5. The experimental results on ranking in Chinese-English OOV translation.

From the experiments above, it can be concluded that the supervised learning significantly outperform the conventional ranking strategies.

5.6 Analysis and Discussion

Through analysis about the experimental results in extraction and ranking, it can be observed that the OOV translation quality is highly related to the following aspects.

(1) The translation results are related to the search engine used, especially for some specific OOV terms. For example, given a query OOV term "两岸三通", the mining result based on Google in China is "*three direct links*", while some meaningless information is mined by the other engines like Live Trans.

(2) Some terms are conventional terminologies and cannot be translated literally. For example, "*woman pace-setter*", a proper name with the particular Chinese characteristic, should be translated into "三八红旗手", rather than "女子的步伐" or "制定".

(3) The proposed model is sensitive to the notability degree of OOV term. For famous person name and book title, the translation performance is very promising. However, for other OOV terms with lower notability, such as "贝尔曼来" and "兰红光", the correct translation cannot even be retrieved by search engine.

(4) Word Sense Disambiguation (WSD) should be added to improve the whole translation performance. Although most of OOVs have unique semantic definition, there are still a few OOVs with ambiguity. For example, "*Rice*" can either be a person name or a kind of food. Another example is "*AARP*", which also has two kinds of meaning, that is, "美国退休者协会" and "地址解析协议".

6 Conclusions and Future Work

In this paper, the proposed model improves the acquirement ability for OOV translation through Web mining and solves the translation pair evaluation problem in a novel way by introducing supervised learning in translation ranking. In addition, it is very significant to apply the key techniques in traditional machine translation into OOV translation, such as OOV recognition, statistical machine learning, alignment of sentence and phoneme, and WSD. The merits of these techniques should be integrated. All these aspects above will become the research focus in our future work.

Acknowledgments

This paper is supported by National Natural Science Foundation of China (No. 60773124), National Science and Technology Pillar Program of China (No. 2007BAH09B03) and Shanghai Municipal R&D Foundation (No. 08dz1500109). Yang Wang is the corresponding author.

References

Chun-Jen Lee, Jason S. Chang, and Jyh-Shing R. Jang. 2006. *Alignment of Bilingual Named Entities in Parallel Corpora Using Statistical Models and Multiple Knowledge Sources.* ACM Transactions on Asian Language Processing, 5(2):121-145.

Gaolin Fang, Hao Yu, and Fumihito Nishino. 2006. *Chinese-English Term Translation Mining Based on Semantic Prediction.* In Proceedings of the COLING/ACL on Main Conference Poster Sessions, pp.199-206.

Jenq-Haur Wang, Jei-Wen Teng, Pu-Jen Cheng, Wen-Hsiang Lu, and Lee-Feng Chien. 2004. *Translating Unknown Cross-Lingual Queries in Digital Libraries Using a Web-based Approach.* In Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, pp.108-116.

Jian-Cheng Wu and Jason S. Chang. 2007. *Learning to Find English to Chinese Transliterations on the Web.* In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp.996-1004.

L. F. Chien. 1997. *PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval.* In Proceedings of SIGIR'97, pp.50-58.

Wen-Hsiang Lu and Lee-Feng Chien. 2004. *Anchor Text Mining for Translation of Web Queries*: *A Transitive Translation Approach.* ACM Transactions on Information Systems, 22(2): 242-269.

Ying Zhang and Phil Vines. 2004. *Detection and Translation of OOV Terms Prior to Query Time.* In Proceedings of SIGIR'04, pp.524-525.

Yunbo Cao, Jun Xu, Tie-Yan LIU, Hang Li, Yalou HUANG, and Hsiao-Wuen HON. 2006. *Adapting Ranking SVM to Document Retrieval.* In Proceedings of SIGIR'06, pp.186-193.