Fusion of Multiple Features and Ranking SVM for Web-based English-Chinese OOV Term Translation

Yuejie Zhang, Yang Wang, Lei Cen, Yanxia Su, Cheng Jin, Xiangyang Xue
School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
{yjzhang,072021176,082024072, 09210240074,jc,xyxue}@fudan.edu.cn

Jianping Fan
Department of Computer Science, The University of North Carolina at Charlotte
jfan@uncc.edu

Abstract

This paper focuses on the Web-based English-Chinese OOV term translation pattern, and emphasizes particularly on the translation selection strategy based on the fusion of multiple features and the ranking mechanism based on Ranking Support Vector Machine (Ranking SVM). By utilizing the CoNLL2003 corpus for the English Named Entity Recognition (NER) task and selected new terms, the experiments based on different data sources show the consistent results. Our OOV term translation model can "*filter*" the most possible translation candidates with better ability. From the experimental results for combining our OOV term translation model with English-Chinese Cross-Language Information Retrieval (CLIR) on the data sets of Text Retrieval Evaluation Conference (TREC), it can be found that the obvious performance improvement for both query translation and retrieval can also be obtained.

1 Introduction

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

With the increasing growth of Web information which includes multilingual hypertext resources with abundant topics, it appears that Web information can mitigate the problem of the restricted OOV term translation accuracy (Lu and Chien, 2002). However, how to select the correct translations from Web information and locate the appropriate translation resources rapidly is still the main goal for OOV term translation. Hence, finding the effective feature representation and the optimal ranking pattern for translation candidates is the core part for the Web-based OOV term translation.

This paper focuses on the Web-based English-Chinese OOV term translation pattern, and emphasizes particularly on the translation selection strategy based on the fusion of multiple features and the translation ranking mechanism based on Ranking Support Vector Machine (Ranking SVM). By utilizing the CoNLL2003 corpus for the English Named Entity Recognition (NER) task and manually selected new terms in various fields, the established OOV term translation model can "*filter*" the most possible translation candidates with better ability. This paper also attempts to apply the OOV term translation mechanism above in English-Chinese CLIR. It can be observed from the experimental results on the data sets of Text Retrieval Evaluation Conference (TREC) that the obvious performance improvement for query translation can be obtained, which is very beneficial to CLIR and can improve the whole retrieval performance.

2 Related Work

At present, the methods for OOV term translation have changed from the basic pattern based on bilingual dictionary, transliteration or parallel corpus to the intermediate pattern based on comparable corpus (Lee et al., 2006; Shao and Ng, 2004; Virga and Khudanpur, 2003), and

Coling 2010: Poster Volume, pages 1435–1443,
Beijing, August 2010

then become a new pattern based on Web mining (Fang et al., 2006; Sproat et al., 2006).

In recent years, many researchers have utilized Web to find the translation candidates on webpages (Wu and Chang, 2007). Al-Onaizan and Knight (2002) used Web statistics information to validate the translation candidates generated by language model, and obtained the accuracy of 72.6% in Arabic-English OOV word translation. Lu and Chien (2004) utilized the statistics information about the anchor texts in Web search results to recognize the translation candidates, and got the accuracy of 63.6% in English-Chinese title query term translation. Zhang and Vines (2004) extracted the translation candidates for OOV query terms in CLIR from Web, and improved the performance of English-Chinese/Chinese-English CLIR to some extent. Zhang et al. (2005) searched the translation candidates by using cross-language query expansion and Web, and obtained the Top-1 accuracy of 81.0% in Chinese-English OOV word translation. Chen and Chen (2006) used the combination of Web statistics and the vocabulary, and acquired the Top-1 accuracy of 87.6% in Chinese-English OOV word translation. Jiang et al. (2007) utilized the combination of Web mining, transliteration and ranking based on Maximum Entropy (ME), only focused on English-Chinese person name translation and got the Top-1 accuracy of 47.5%.

Although the methods above can improve the translation performance for OOV term to a certain degree, there are still three common problems in the OOV term translation based on Web mining. (1) **Chinese key term extraction pattern from Web documents is over complex and the complexity is always higher.** Because of the inherent property of having no segmentation delimitation in Chinese, it's very difficult for English-Chinese OOV term translation to extract Chinese key terms from Web documents. The cost for the extraction computation is generally overlarge (Wang et al., 2004; Zhang and Vines, 2004). (2) **The feature information for the evaluation of translation candidates is not enough and comprehensive.** Most of OOV term translation methods implement the evaluation for candidates through mining simple local and Boolean features, that is, inherent features in candidates and their surrounding context features. However, if only a certain Web document that an OOV term appears is explored, the global information contained in the whole Web document set will be ignored, and the inconsistency and polysemy of candidates cannot be considered. (3) **The relevance measurement for translation pairs is very simple, or the computation cost is too high.** For ranking candidates, most of OOV term translation approaches adopt the simple combination computation of the feature values used, or get assessment based on classification models. Hence, the feature weights are determined according to the corresponding induction and suitable for some specific fields, but cannot guarantee the accuracy of the final translation ranking results. However, the Ranking SVM model can effectively express multiple ranking constraints, and has better universality and applicability (Cao et al., 2006; Joachimes, 2002; Vapnik, 1995).

3 Our Solutions

To support more precise English-Chinese OOV term translation, we establish a multiple-feature-based translation pattern based on Web mining and Ranking SVM. On the one hand, a Chinese key term extraction strategy is built on the simplified extraction computation for PAT-Tree, in which the optimization processing for the confidence of word building is improved to a certain extent. On the other hand, translation candidates are chosen by the fusion of multiple features. The representation forms of local, global and Boolean feature are constructed under the consideration of the complex characteristics of English/Chinese OOV term and Web information. Moreover, for the relevance measurement between an OOV term to be translated and its translation candidates, the supervised learning based on Ranking SVM is introduced to rank candidates precisely.

At first, given an OOV term to be translated as a query, it is input into the Google search engine to acquire the returned webpage snippet set. Next, Chinese key terms are extracted from the PAT-Tree built on the snippet set to determine the translation candidates. Subsequently, local, global and Boolean features are extracted from the candidates based on the fusion of multiple features. Finally, the candidates are filtered and ranked through the supervised learning based on Ranking SVM.

4 Chinese Key Term Extraction

In Web mining of English-Chinese OOV term translation, an important problem is to extract the target translation candidates from the returned Chinese Web documents, which can be considered as a key term extraction task.

The PAT-Tree structure is an efficient indexing method in both IR and Information Extraction (IE) domains (Chien, 1997; Gonnet et al., 1992). Its superior feature is the Semi Infinite String, which can store all the strings from the whole corpus (i.e., the returned snippet set in this paper) in a binary tree. The branch node indicates the search direction and the leaf node stores the index and frequency for a string.

Generally, a Chinese character corresponds to a binary-coded form with 2 bytes (16 bits). Chinese strings can be transformed into binary strings. There is an ending tag for each string and its binary form is "00000000". Take "中文信息抽取" (*Chinese IE*) and "信息检索" (*IR*) as an example, the binary strings for them are described in Figure 1. Thus a PAT-Tree can be built based on these strings, as shown in Figure 2. The branch node stands for the comparison bit (Comp-bit), which represents the position of different bit in binary strings. Some binary strings have the value of 0 in such a bit and are classified into the left branch, while others have 1 and turn to the right branch.

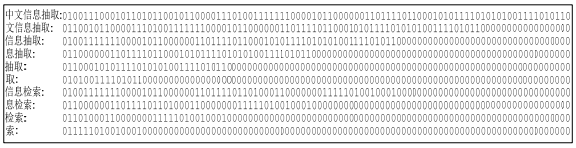

Figure 1. Binary string representation instantiation.

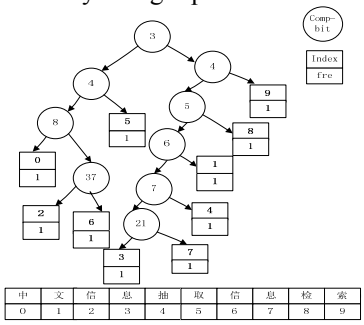

Figure 2. PAT-Tree Instantiation for Figure 1.

In the extraction process, the PAT-tree is traversed first, and the branch nodes with the Comp-bit values larger than 32 are selected. This is because the minimum length of a Chinese common string is 2 characters and each has 16 binary bits. Next, the frequency values of both two child nodes are added as the frequency of the common string (i.e., the parent branch node). At last, the common strings with the frequency values larger than 2 are extracted as the key terms. For the PAT-Tree in Figure 2, there is a branch node with the Comp-bit value of 37, which indicates that at least the prefixes of two strings contain two identical characters. It can be known from the leaf nodes that two strings are "信息抽取" (*IE*) and "信息检索" (*IR*). Hence, the prefix substring "信息" (*information*) with the frequency of 2 is extracted as the common string. Thus the key terms with the arbitrary lengths and frequency values can be retrieved from the built PAT-Tree.

However, with the common strings being extracted, large amounts of noisy terms and fragments are also extracted. To filter noisy fragments, Wang et al. (2004) used SPDCD and the Local-Maxima algorithm, but the computation cost was too expensive. Therefore, the simplified filtering manner is adopted here:

$$\alpha(c_i \cdots c_j) = \frac{f(c_i \cdots c_j) - f(c_1 \cdots c_n)}{f(c_i \cdots c_j)} \qquad (1)$$

where $c_1 \ldots c_n$ is a n-gram that contains the substring $c_i \ldots c_j$; $c_i \ldots c_j$ is the n-1-gram to be estimated, i.e., $c_i \ldots c_j = c_1 \ldots c_{n-1}$ or $c_i \ldots c_j = c_2 \ldots c_n$; $f(\ )$ denotes the string frequency; α represents the cohesion factor of the n-1-gram string, that is, the ability of independent word building. The closer to 1 the value of α is, the more possible meaningful key term $c_i \ldots c_j$ is.

5 Multiple Feature Representation

Local Feature (LF) is constructed based on neighboring tokens and the token itself. There are two types of contextual information to be considered when extracting LFs, namely internal lexical and external contextual information.

(1) **Term length** (*Len*) – Aims to consider the length of the translation candidate.

(2) **Phonetic Value** (*PV*) – Aims to investigate the phonetic similarity between an OOV term and its translation candidates. Because the associated syllabification representations can often be found between Chinese and English syllables with fewer ambiguities, the syllabification has become an effective channel in phonetic feature expression. *PV* means that for measuring the edit distance similarity between the syllabification sequences of an OOV term

and its candidates, the processing is executed according to the specific linguistic rules.

$$PV(S_{OOV}, T_{OOV}) = 1 - \frac{EditDist(S_{OOV}', T_{OOV}')}{Len(S_{OOV}') + Len(T_{OOV}')} \quad (2)$$

where S_{OOV} and T_{OOV} denote the OOV term in the source language and its translation candidate in the target language respectively, S_{OOV}' and T_{OOV}' are the character strings after the syllabification and removing the vowels, *EditDist*(,) indicates the edit distance between two strings, and *Len*( ) is the string length.

(3) **Length Ratio of OOV Term and Its Translation Candidate** (*LR*) – Aims to explore the composition possibility that the extracted key term can be regarded as the translation for an OOV term. An OOV term and its translation should have the similar length, so the *LR* value is close to 1 as possible. A Chinese term is segmented into significant pieces first, and the number of pieces is taken as its length. For example, "非典型肺炎" (*SARS*) is segmented into "非" (*non*), "典型" (*typical*) and "肺炎" (*pneumonia*), and its length is 3. For an English term, the number of words is counted as the length. If there is only one word composed of capital letters, its length is defined as the number of letters, e.g., "*SARS*" has the length of 4. Thus the *LR* value of "*SARS*" and its candidate "非典型肺炎" is 4/3=1.3.

(4) **Phonetic and Semantic Integration Feature** (*P&S_IF*) – Aims to consider the phonetic information and senses of an OOV term and its candidates synthetically. It is set up for multi-word OOV terms, especially for NEs and new terms. Each constituent can be translated by the phonetic information or senses.

$$P\&S\_IF(S_{OOV}, T_{OOV}) = \frac{LScore(S_{OOV}, T_{OOV}) + PV(S_{OOV}'', T_{OOV}'')}{LScore(S_{OOV}, T_{OOV}) + 1} \quad (3)$$

where *LScore*(,) is the matching word number of non-transliteration words in S_{OOV} and T_{OOV}, while S_{OOV}'' and T_{OOV}'' are the remaining strings of S_{OOV} and T_{OOV} after computing *LScore*. For example, given S_{OOV} "*Capitoline Museum*" and its T_{OOV} "卡比多里尼博物馆" (*Capitoline Museum*), the non-transliteration words "*Museum*" and "博物馆" (*museum*) are matched, then *LScore*(S_{OOV}, T_{OOV})=1; the *PV* value between the remaining strings "*Capitoline*" and "卡比多里尼" (*Capitoline*) is 0.8, so the final *P&S_IF* value is 1.8/2=0.9.

Global Feature (GF) is extracted from other occurrences of the same or similar tokens in the Web document set. The common case in the Web-based OOV term translation is that the translation candidates in the previous parts of Web documents will often occur with the same or similar forms in the latter parts. The contextual information from the same and other Web documents may play an important role in determining the final translation. To utilize such global information, GFs are constructed based on the characteristics of Web documents.

(1) **Global Term Frequency** (*G_Freq*) – Aims to utilize the frequency information that an OOV term and its translation candidates appear in the Web document set. It is always the most important feature and includes four parameters. $\boldsymbol{Freq_{Soov}}$ denotes the frequency of S_{OOV} in all the returned webpage snippets. $\boldsymbol{TF_{Toov}}$ indicates the number of T_{OOV}s in all the snippets. $\boldsymbol{DF_{Toov}}$ represents the number of snippets that contain T_{OOV}. ***CO_Freq*** means the number of snippets that contain both S_{OOV} and T_{OOV}, i.e, co-occurrence frequency.

(2) **Chi-Square** (χ^2) **Feature Value** (*CV*) – Aims to evaluate the semantic similarity between an OOV term and its translation candidates by their occurrence in Web documents.

$$CV_{\chi^2}(S_{OOV}, T_{OOV}) = \frac{N \times (a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)} \quad (4)$$

where a is the number of snippets that contain both S_{OOV} and T_{OOV}, b is the number of snippets that contain S_{OOV} but do not contain T_{OOV}, c is the number of snippets that do not contain S_{OOV} but contain T_{OOV}, d is the number of snippets that do not contain neither of S_{OOV} and T_{OOV}, and $N=a+b+c+d$.

(3) **Co-occurrence Distance** (*CO_Dist*) – Aims to investigate the distance between an OOV term and its candidates in Web documents. This distance is often very closer.

For each snippet that contains both S_{OOV} and T_{OOV}, three positions are considered, that is, the first position that S_{OOV} and T_{OOV} appear ($p1$), the second position ($p2$) and the last one ($p3$). In the following snippet, S_{OOV} is "*AARP*" and T_{OOV} is "美国退休者协会" (*America Association of Retired Persons*, *AARP*).

拿什么创造《AARP杂志》的成功--传媒--人民网
2008年10月18日 ... 它是一个协会杂志，隶属于美国退休者协会（AARP）。《AARP杂志》自称是"世界上发行量 ... 笔者认为，《AARP杂志》的成功，很大程度上得益于以下几点。 ...
media.people.com.cn/GB/22114/45733/136325/8192992.html - 38k - 网页快照 - 类似网页

$p1_{SOOV}$=6, $p2_{SOOV}$=62, $p3_{SOOV}$=97;
$p1_{TOOV}$=54, $p2_{TOOV}$=-1, $p3_{TOOV}$=54.

The position is indexed from 0 and $p2_{TOOV}$=-1 means only one candidate exists in the snippet. Then the nearest position pair $p2_{SOOV}$ and $p1_{TOOV}$ can be found for this example. The distance *Dist* between S_{OOV} and T_{OOV} is computed as:

$$Dist(S_{OOV}, T_{OOV}) = \begin{cases} pi_{S_{OOV}} - pj_{T_{OOV}} - Len(T_{OOV}), & pi_{S_{OOV}} > pj_{T_{OOV}} \\ pj_{T_{OOV}} - pi_{S_{OOV}} - Len(SOOV), & pi_{S_{OOV}} < pj_{T_{OOV}} \end{cases} \quad (5)$$

Given the example above, $Dist=p2_{SOOV}-p1_{TOOV}-7=62-54-7=1$, that is, S_{OOV} and T_{OOV} are a left bracket '(' apart. Finally, the average distance *CO-Dist* in the snippet set can be computed as:

$$CO\_Dist(S_{OOV}, T_{OOV}) = AVG\_Dist(S_{OOV}, T_{OOV}) = \frac{Sum(Dist)}{CO\_Freq(S_{OOV}, T_{OOV})} \quad (6)$$

where *Sum*( ) is the sum of *Dist* in each snippet. (4) **Rank Value** (*RV*) – Aims to consider the rank for translation candidates in the Web document set. It includes five parameters. ***Top_Rank*** (*T_Rank*) is the rank of the snippet that first contains T_{OOV} and given by the search engine. ***Average_Rank*** (*A_Rank*) is the average position of T_{OOV} in the returned snippets.

$$A\_Rank(T_{OOV}) = \frac{Sum(Rank)}{DF_{T_{OOV}}(T_{OOV})} \quad (7)$$

where *Sum*( ) denotes the rank sum of each snippet. ***Simple_Rank*** (*S_Rank*) is computed as $S\_Rank(T_{OOV})=TF_{TOOV}(T_{OOV})*Len(T_{OOV})$, which aims at investigating the impact of the frequency and length of T_{OOV} on ranking. ***R_Rank*** is utilized as a comparison basis.

$$R\_Rank(T_{OOV}) = \beta \times \frac{|T_{OOV}|}{MAX\_WL} + (1-\beta) \times \frac{TF_{T_{OOV}}(T_{OOV})}{Freq_{S_{OOV}}(S_{OOV})} \quad (8)$$

where β is set as 0.25 empirically, $|T_{OOV}|$ is the length of T_{OOV}, and *MAX_WL* denotes the maximum length of candidate terms. ***DF_Rank*** (*D_Rank*) is similar to *S_Rank* and computed as $D\_Rank(T_{OOV})=DF_{TOOV}(T_{OOV})*Len(T_{OOV})$.

Boolean Feature (BF) is a binary feature and equivalent to a heuristic rule designed for the particular relationship between an OOV term and its translation candidates. BFs are used to explore the different occurrence forms with higher possibility for the translation candidates in Web documents. (1) **Position Distance with OOV Term** (*PD_S*$_{OOV}$) – If T_{OOV} occurs close to S_{OOV} (within 10 characters), then this feature is set as 1, else -1. (2) **Neighbor Relationship with OOV Term** (*NR_S*$_{OOV}$) – If T_{OOV} occurs prior or next to S_{OOV}, then this feature is set as 1. (3) **Bracket Neighbor Relationship with OOV Term** (*BNR_S*$_{OOV}$) – If T_{OOV} locates prior or next to S_{OOV} and occurs with the form "T_{OOV} (S_{OOV})" or "S_{OOV} (T_{OOV})", then this feature is set as 1. (4) **Special Mark Word** (*SMW*) – This is an intuitive feature. Within a certain co-occurrence distance (usually less than 10 characters) between an OOV term and its candidates, if there is such a term like "全称" (*full name*), "叫" (*be named as*), "译为" (*be translated as* …), "名称" (*name*), or "(或/又)称为" ((*or*/*also*) *be called as* …), or within 5 characters if there are some punctuations like "( )", "[ ]" and "（）", then this feature is set as 1.

6 Ranking based on Ranking SVM

For the OOV term translation based on Web mining, another difficulty is how to evaluate the relevance between an OOV term and its translation candidates, that is, how to rank the translation candidates from "*best*" to "*worst*".

The candidate ranking can be regarded as a binary classification problem. However, usually only highly related fragments of OOV terms can be found, rather than their correct translations. Instead of regarding the candidate ranking as binary classification, it is solved as an Ordinal Regression problem. Ranking SVM maps different objects into a certain kind of order relation. The key is modeling the judgements for user's preferences, and then the constraint relations for ranking can be derived (Herbrich et al., 1999; Xu et al., 2005).

For a given OOV term S_{OOV}, if there are two translation candidates T_{OOVi} and T_{OOVj}, the preference judgement can be formulated as $T_{OOVi} \succ_{SOOV} T_{OOVj}$. Thus more training samples are constructed, which contain multiple constraint features. The preference judgement can be transformed into the feature function as:

$$f(w, T_{OOVi}, S_{OOV}) \succ_{SOOV} f(w, T_{OOVj}, S_{OOV}) \quad (9)$$

where w is a parameter and represented as a n-dimensional vector $w=\{w_1, w_2, \ldots, w_n\}$. This feature function can also be expressed as:

$$f(w, T_{OOV}, S_{OOV}) = \sum_{k=1}^{p} w_k LF_k(T_{OOV}, S_{OOV}) + \sum_{l=p+1}^{q} w_l GF_l(T_{OOV}, S_{OOV}) + \sum_{m=q+1}^{n} w_m BF_m(T_{OOV}, S_{OOV}) \quad (10)$$

where LF_k(,), GF_l(,) and BF_m(,) are the local, global and Boolean feature representation respectively. These three kinds of feature representation are incorporated as a whole and represented as a feature function family with the multi-dimensional feature vector in (11).

$$f(w, T_{OOV}, S_{OOV}) = w \cdot h(T_{OOV}, S_{OOV}) \quad (11)$$

That is the ranking results for candidates. Thus the relevance for each feature vector x (translation candidate) containing a group of features can be evaluated through Ranking SVM.

7 Experiment and Analysis

7.1 Data Set and Evaluation Metrics

For the performance evaluation, 4,593 English NEs are selected from the English corpus of the NER task in CoNLL2003. The test set contains 446 Person Names (PRNs), 329 Location Names (LCNs) and 455 Organization Names (OGNs), and the remaining is taken as the training set (including 1,137 PRNs, 1,152 LCNs and 1,074 OGNs) through manually tagging. Additionally, 300 English new terms are chosen randomly from 9 categories, including movie name, book title, brand name, terminology, idiom, rare animal name, rare PRN and OGN. Such terms are used to investigate the generalization ability of our model.

Top-N-Inclusion-Rate is used as a measurement for the translation performance. For a set of OOV terms to be translated, its *Top-N-Inclusion-Rate* is defined as the percentage of the OOV terms whose translations could be found in the first N extracted translations.

7.2 Experiment on Parameter Setting

For Chinese key term extraction, the test on the threshold α is performed. As shown in Figure 3, when the lower bound of α is set as 0.4, the best performance can be achieved.

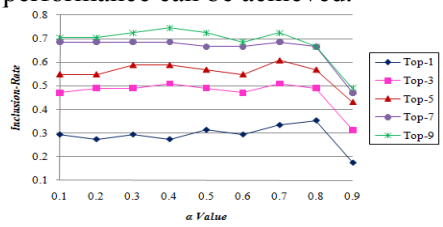

Figure 3. Results for α value setting.

To get the most relevant candidates into top-10 before the final ranking, an initial ranking test is performed on *S_Rank*, *R_Rank* and *D_Rank*. It can be seen from Figure 4 that *D_Rank* exhibits the better performance.

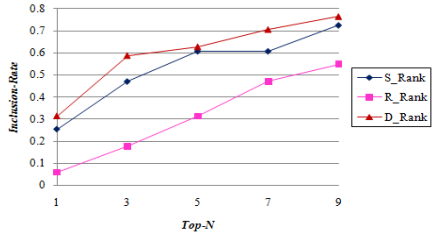

Figure 4. Results for initial ranking manner.

To find how many returned webpage snippets are suitable for the translation acquisition, the test on the snippet number is performed. As shown in Figure 5, the best performance can be obtained by using 200 snippets.

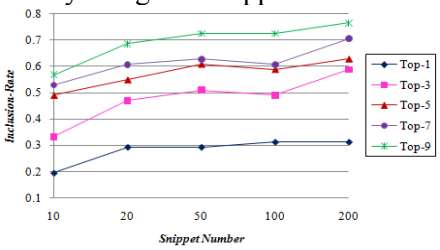

Figure 5. Results for webpage snippet number.

7.3 Experiment on Multiple Feature Fusion

To verify the effectiveness for multiple feature fusion, the test on the feature combination for OOV term translation is implemented. As shown in Table 1, the highest accuracy (the percentage of the correct translations in all the extracted translations) of 83.1367% can be acquired by using all the features.

Feature				Accuracy	Reduction
All Features				83.1367%	—
Numerical Feature	*Local Numerical Feature*	*-Len*		81.7355%	-1.4012%
		-PV		77.4494%	-5.6873%
		-LR		81.4231%	-1.7136%
		-P&S_IF		79.9002%	-3.2365%
	Global Numerical Feature	*Global Frequency*	*-TF_{Toov}*	82.9877%	-0.1490%
			-DF_{Toov}	83.2112%	+0.0745%
			-CO_Freq	83.0870%	-0.0497%
		-CV		82.3125%	-0.8242%
		-CO_Dist		81.8577%	-1.2790%
		RV	*-T_Rank*	83.0125%	-0.1242%
Boolean Feature		*-PD_S_{OOV}*		82.1806%	-0.9561%
		-NR_S_{OOV}		82.2923%	-0.8444%
		-BNR_S_{OOV}		80.7525%	-2.3842%
		-SMW		83.1740%	+0.0373%

Table 1. Results for feature combination.

In Table 1, '-' before the specific feature denotes that the OOV term is translated by combining all the other features except this feature; "*Reduction*" represents the difference value between the translation accuracy obtained by using all the features and that by removing a specific feature. The positive "*Reduction*" indicates that the accuracy is improved after removing a specific feature, while the negative shows the accuracy is decreased.

It can be seen from Table 1 that for mining the translations for OOV terms, the most important three features are *PV*, *P&S_IF* and *BNR*, then *LR*, *Len* and *CO_Dist*. As for the frequency feature, its contribution is limited, because many translation candidates with higher *PV* or *P&S_IF* values are the terms with low frequency. It shows that *PV* and *P&S_IF* play a very crucial role in mining the translation candidates with low frequency. In addition,

the contribution degree of *CV* is also positive. However, when training based on only the features that are beneficial to the whole translation performance, the best translation accuracy is 83.1243%, which is worse than that by combining all the features. From a view of the effect of the single feature on the whole translation performance, some features may have slightly negative impact. Nevertheless, through combining all the features, the multiple feature fusion mechanism can indeed efficiently improve the translation accuracy.

7.4 Experiment on OOV Term Translation

Some translation examples based on different ranking patterns are given in Table 2, in which the score represents the correlation degree between the translation pair. The closer to -1 the score is, the more irrelevant the translation pair is; while the closer to 3 the score is, the more relevant the translation pair is.

PRN -- "*Santamaria*"		
Candidates (Top-5)	**SVM Score**	**Ranking SVM Score**
桑塔马利雅	1.1746	3.17754
辛达马利亚	0.7087	2.81014
桑塔玛利亚	0.9326	2.68914
圣何塞	0.2879	2.26468
蒙哥山塔马利亚	0.2051	2.1525
LCN -- "*Gettysburg National Military Park*"		
Candidates (Top-5)	**SVM Score**	**Ranking SVM Score**
葛底斯堡国家军事公园	0.7500	2.4998
堡国家军事公园	0.6666	2.4159
国家军事公园	0.3973	1.8539
盖茨堡国家军事公园	0.2877	1.5172
在葛底斯堡建立了国家军事公园	-0.3407	0.8019
OGN -- "*Federal Reserve Board*"		
Candidates (Top-5)	**SVM Score**	**Ranking SVM Score**
美国联准会	0.9784	2.7435
美国联邦储备委员会	0.9483	2.7314
美国联邦准备制度	0.5387	2.7178
联邦储备金监察小组	1.2031	2.6684
联邦准备理事会	0.7425	2.6003

Table 2. OOV term translation examples.

Furthermore, Jiang et al. (2007) utilized the combination of Web mining, transliteration and ME-based ranking to implement English-Chinese PRN translation, which is very similar to our approach. To make a contrast with it, we accomplished this method on the same data set. The comparison results are shown in Table 3.

Ranking Pattern	**Category**	**Top-1**	**Top-2**	**Top-3**
***based on SVM* (*Multiple Features*)**	***PRN***	64.44%	85.07%	91.42%
	LCN	53.93%	73.33%	81.82%
	OGN	49.68%	70.70%	82.16%
	All	**56.10%**	**76.59%**	**85.45%**
***based on Ranking-SVM* (*Multiple Features*)**	***PRN***	77.14%	89.20%	93.96%
	LCN	64.24%	75.15%	85.45%
	OGN	63.05%	79.61%	89.17%
	All	**68.46%**	**81.87%**	**89.92%**
[Jiang et al., 2007] *based on ME* ($PV+CV+NR\_S_{OOV}+BNR\_S_{OOV}$)	***PRN* (Only)**	49.07%	57.33%	60.43%

Table 3. Performance comparison results.

From the experimental results above, it can be concluded that the ranking based on the supervised learning significantly outperforms the conventional ranking strategies, and Ranking SVM is superior to SVM and ME for translation candidate ranking. From the contrast between our model and Jiang's method, it can be found that our approach is superior to Jiang's and the better performance can be achieved based on the fusion of multiple features proposed in this paper. Meanwhile, it can also be observed from Table 3 that the performance for LCN and OGN translation is better, while the best performance is obtained for PRN translation. It shows that our translation model is sensitive to the category and the popularity degree of OOV term to some extent.

In order to test the translation performance for the other kinds of English OOV term, another test is performed based on the OOV new terms selected randomly from 9 categories. The experimental results are shown in Table 4.

Top-N-Inclusion-Rate	**Top-1**	**Top-3**	**Top-5**	**Top-7**	**Top-9**
Other OOV Terms	49.41%	71.02%	72.46%	81.51%	84.30%

Table 4. Results for other OOV terms.

Furthermore, the translations for some OOV terms based on different translation manners are compared, including our proposed model, Google Translate and the Live Trans translation model developed by WKD Lab at National Taiwan University, as shown in Table 5.

OOV Terms	**Translation from Our Model**	**Translation from Google Translate**	**Translation from Live Trans**
Forrest Gump	阿甘正传/电影	阿甘正传	阿甘正传/亚伦席维斯崔
Estee Lauder	雅诗兰黛/化妆品	雅诗兰黛	雅诗兰黛/香水/化妆品
Arteriosclerosis	动脉硬化	动脉粥样硬化	心脏/动脉硬化
Woman Pace-Setter	三八红旗手	女子的步伐/制定	三八红旗手
Dream of the Red Mansion	红楼/红楼梦	红楼梦	红楼梦/文章书目
SARS	非典型肺炎/非典	严重急性呼吸系统综合症	病毒/非典型肺炎
NASA	美国宇航局	美国航天局	美国太空总署

Table 5. Comparison for different translation manners.

The results above demonstrate that our model can be applicable to all kinds of OOV terms and has better translation performance.

7.5 Experiment on English-Chinese CLIR

To explore the applicability and usefulness of our OOV term translation model in English-Chinese CLIR, four CLIR runs based on *long query* (terms in both title and description fields) and *short query* (only terms in the title field) are carried out on the English topic set (25 topics) and Chinese corpus (127,938 documents) from TREC-9. (1) *E-C_LongCLIR*1 – using long query and the bilingual-dictionary-based query translation; (2) *E-C_LongCLIR*2 – using long query, the bilingual-dictionary-based

query translation and our OOV term translation; (3) *E-C_ShortCLIR*1 – using short query and the bilingual-dictionary-based query translation; (4) *E-C_ShortCLIR*2 – using short query, the bilingual-dictionary-based query translation and our OOV term translation. The Precision-Recall curves and Median Average Precision (MAP) values are shown in Figure 6.

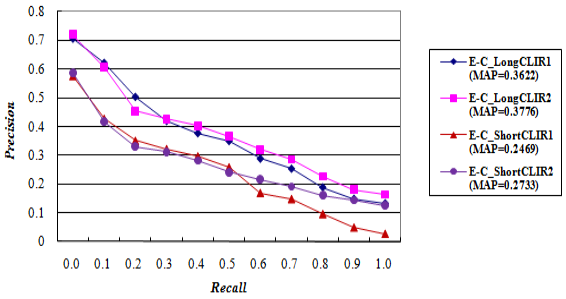

Figure 6. Results for English-Chinese CLIR combining our OOV term translation model.

It can be seen from Figure 6 that the best run is *E-C_LongCLIR*2, and its results exceed those by another run *E-C_LongCLIR*1 based on long query. By adopting both query translation based on bilingual dictionary and OOV term translation, the English-Chinese CLIR for long query has gained the significant improvement on the whole retrieval performance. Compared with the traditional query translation based on bilingual dictionary, such a combination manner is exactly a better way for query translation from the source language to the target language. Additionally, through comparing the results for the other two runs *E-C_ShortCLIR*1 and *E-C_ShortCLIR*2 based on short query, it can also be further confirmed that our OOV term translation mechanism can also support CLIR for short query effectively.

7.6 Analysis and Discussion

Through analyzing the results for translation extraction and ranking, it can be found that the translation quality is highly related to the following aspects. (1) **The translation results are associated with the search engine used, especially for some specific OOV terms.** For example, given an OOV term "*Cross-Strait Three-links*", the mining result based on Google in China is "两岸大三通", while some meaningless information is mined by Live Trans. (2) **Some terms are conventional terminologies and cannot be translated literally.** For example, "*Woman Pace-Setter*", a proper noun with the Chinese characteristic, should be translated into "三八红旗手", rather than "女子的步伐" (*women's pace*) or "制定" (*establishment*) given by Google Translate. (3) **The proposed model is sensitive to the notability degree of OOV term.** This phenomenon is the main reason why there is obvious difference among the translation performance for PRN, LCN and OGN. (4) **There is a "*fragment effect*" in PAT-Tree-based Chinese key term extraction.** The fragments of Chinese terms have become the main noisy data. Such a problem should be solved by setting the specific threshold for additional features like heuristic rules and occurrence distance. (5) **Word Sense Disambiguation (WSD) should be added to improve the translation performance.** Although most of OOV terms have a unique semantic definition, there are still a few OOV terms with ambiguity, e.g., "*AARP*" (*American Association of Retired Persons* or *AppleTalk Address Resolution Protocol*). (6) **The ranking pattern based on the supervised learning is able to synthesize various feature representations for translation candidates.** Thus the rank for a candidate can be precisely predicted through tagging and training.

8 Conclusions

In this paper, the proposed model improves the acquirement ability for OOV term translation through Web mining, and solves the translation pair selection and evaluation in a novel way by fusing multiple features and introducing the supervised learning based on Ranking SVM. Furthermore, it is significant to apply the key techniques in machine translation into OOV term translation, such as OOV term recognition, statistical machine learning, alignment of sentence and phoneme, and WSD. All these aspects will be our research focus in the future.

Acknowledgements

This work is supported by National Natural Science Fund of China (No. 60773124), Shanghai Natural Science Fund (No. 09ZR1403000), National Science and Technology Pillar Program of China (No. 2007BAH09B03), 973 Program of China (No. 2010CB327906), Shanghai Municipal R&D Foundation (No. 08dz1500109) and Shanghai Key Laboratory of Intelligent Information Processing. Cheng Jin from Fudan University is the corresponding author.

References

Y. Al-Onaizan, K. Knight. 2002. *Translating Named Entities using Monolingual and Bilingual Resources*. In: The 30th Meeting of the Association for Computational Linguistics (ACL 2002), 400-408.

Y.B. Cao, J. Xu, T.Y. Liu, H. Li, Y.L. Huang, and H.W. Hon. 2006. *Adapting Ranking-SVM to Document Retrieval*. In: The 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006), 186-193.

C. Chen, H.H. Chen. 2006. *A High-Accurate Chinese-English NE Backward Translation System Combining Both Lexical Information and Web Statistics*. In: The Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL 2006), 81-88.

L.F. Chien. 1997. *PAT-Tree-based Keyword Extraction for Chinese Information Retrieval*. In: The 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1997), 50-58.

G.L. Fang, H. Yu, and F. Nishino. 2006. *Chinese-English Term Translation Mining Based on Semantic Prediction*. In: The Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL 2006), 199-206.

G.H. Gonnet, R.A. Baeza-Yates, and T. Sinder. 1992. *New Indices for Text: PAT Trees and PAT Arrays*. Information Retrieval Data Structures & Algorithms, 66-82.

R. Herbrich, T. Graepel, and K. Obermayer. 1999. *Support Vector Learning for Ordinal Regression*. In: The 9th International Conference on Neural Networks (ICANN 1999), 97-102.

L. Jiang, M. Zhou, L.F. Chien, and C. Niu. 2007. *Named Entity Translation with Web Mining and Transliteration*. In: The 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), 1629-1634.

T. Joachimes. 2002. *Optimizing Search Engines using Click through Data*. In: The 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2002), 133-142.

C.J. Lee, J.S. Chang, and J.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.

W.H. Lu, L.F. Chien. 2002. *Translation of Web Queries using Anchor Text Mining*. ACM Transactions on Asian Language Information Processing, 1(2):159-172.

W.H. Lu, L.F. Chien. 2004. *Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach*. ACM Transactions on Information Systems, 22(2):242-269.

L. Shao, H.T. Ng. 2004. *Mining New Word Translations from Comparable Corpora*. In: The 20th International Conference on Computational Linguistics (COLING 2004), 618-624.

R. Sproat, T. Tao, and C.X. Zhai. 2006. *Named Entity Transliteration with Comparable Corpora*. In: The Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL 2006), 73-80.

V.N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY.

P. Virga, S. Khudanpur. 2003. *Transliteration of Proper Names in Cross-Language Applications*. In: The 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), 365-366.

J.H. Wang, J.W. Teng, P.J. Cheng, W.H. Lu, and L.F. Chien. 2004. *Translating Unknown Cross-Lingual Queries in Digital Libraries using a Web-based Approach*. In: The Joint Conference on Digital Libraries (JCDL 2004), 108-116.

J.C. Wu, J.S. Chang. 2007. *Learning to Find English to Chinese Transliterations on the Web*. In: The Joint Meeting of the Conference on Empirical Methods in Natural Language Processing and the Conference on Computational Natural Language Learning (EMNLP-CoNLL 2007), 996-1004.

J. Xu, Y.B. Cao, H. Li, and M. Zhao. 2005. *Ranking Definitions with Supervised Learning Methods*. In: The 14th International World Wide Web Conference (WWW 2005), 811-819.

Y. Zhang, P. Vines. 2004. *Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval*. In: The 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), 162-169.

Y. Zhang, P. Vines. 2004. *Detection and Translation of OOV Terms Prior to Query Time*. In: The 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), 524-525.

Y. Zhang, F. Huang, and S. Vogel. 2005. *Mining Translations of OOV Terms from the Web through Cross-Lingual Query Expansion*. In: The 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005), 669-670.