Chinese-English OOV Term Translation with Web Mining, Multiple Feature Fusion and Supervised Learning

Yun Zhao[1], Qinen Zhu[1], Cheng Jin[1], Yuejie Zhang[1], Xuanjing Huang[1], Tao Zhang[2]

[1]School of Computer Science
Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, Shanghai 200433, P. R. China
[2]School of Information Management & Engineering,
Shanghai University of Finance & Economics, Shanghai 200433, P. R. China
{12210240077, 13210240131, jc, yjzhang, xjhuang}@fudan.edu.cn,
taozhang@mail.shfeu.edu.cn

Abstract. This paper focuses on the Web-based Chinese-English Out-of-Vocabulary (OOV) term translation pattern, and emphasizes on the translation selection based on multiple feature fusion and the ranking based on Ranking Support Vector Machine (Ranking SVM). By utilizing the SIGHAN2005 corpus for the Chinese Named Entity Recognition (NER) task and selected new terms, the experiments based on different data sources show the consistent results. From the experimental results for combining our model with Chinese-English Cross-Language Information Retrieval (CLIR) on the data sets of TREC, it can be found that the obvious performance improvements for both query translation and CLIR are obtained.

Keywords: Chinese-English OOV Term Translation, Web mining, multiple feature fusion, supervised learning, Ranking SVM.

1 Introduction

In Cross-Language Information Retrieval (CLIR), 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 [1][5][6][12]. The translation quality of OOV term directly influences the precision of querying multilingual information and OOV term translation has become a challenging issue in CLIR [9][15][17]. 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 [11][13][18]. However, how to select the correct translations from Web and locate the appropriate translation resources rapidly is still the main goal for OOV term translation [14][16][19]. 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.

Many researchers have utilized Web search engines to find translation candidates for Chinese-English OOV term translation [8][10][13]. Zhang et al. [25] extracted the translation candidates for OOV query terms from Web in Chinese-English CLIR, and improved the CLIR performance. Zhang et al. [24] 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. Fang et al. [4] used semantic prediction and query expansion to get the translation candidates, and acquired the Top-3 accuracy of 82.9% in Chinese-English OOV term translation. Chen et al. [3] used the combination of Web statistics and the vocabulary, and acquired the Top-1 accuracy of 87.6% in Chinese-English OOV word translation. Yang et al. [21] utilized the combination of transliteration, Web mining and ranking based on AdaBoost, and got the Top-5 accuracy of 76.35% for Chinese-English backward transliteration. Yang et al. [22] utilized heuristic Web mining and asymmetric alignment, and got the Top-1 accuracy of 48.71% in Chinese-English organization name translation. Yang et al. [23] combined Web mining and ranking by SVM and Ranking SVM, and obtained the Top-1 accuracy of 65.75% in Chinese-English organization name translation.

Unfortunately, there are still three common problems in Chinese-English OOV term translation based on Web mining. (1) **The noises in English translation candidates cannot be processed appropriately.** Although there does not exist the issue of word segmentation in English key term extraction, many noises may be introduced into the candidates extracted from Web documents. However, such noises are often simply processed, or even without any processing. (2) **The feature information for the evaluation of translation candidates is not enough and comprehensive.** Most methods implement the evaluation for candidates through mining simple local and Boolean 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 is ignored, and the inconsistency and polysemy of candidates cannot be considered. (3) **The relevance measurement for translation pairs is simple, or the computation cost is too high.** For ranking candidates, most approaches adopt the simple combination computation of feature values, or get assessment based on classification models. The feature weights are determined according to the general induction and suitable for specific fields, and cannot guarantee the accuracy for ranking. The Ranking SVM model can effectively express multiple ranking constraints, and has better universality and applicability [2][20].

To support more precise Chinese-English OOV term translation, we establish a multiple-feature-based translation pattern based on Web mining and Ranking SVM. An English key term extraction mechanism is built on the simplified selection, and then the emphasis is put on the noise filtering. Heuristic rules summarized from translation candidates are used to remove insignificant noises, and Information Entropy is introduced to further discard meaningless substrings. 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 for the characteristics of Chinese/English OOV term and Web information. For the relevance measurement between an OOV term and its translation candidates, the supervised learning based on Ranking SVM is utilized to rank candidates accurately. By utilizing

the SIGHAN2005 corpus for the Chinese Named Entity Recognition (NER) task and manually selected new terms in various fields, our model can "*filter*" the most possible translation candidates with better ability. This paper also attempts to apply our model in Chinese-English CLIR. It can be observed from the experimental results on the data sets of TREC that the obvious improvement for query translation is obtained.

2 English Key Term Extraction

In Web mining of OOV term translation, a crucial problem is to select the translation candidates from the returned Web documents, that is, the key term extraction task. The **Initial Extraction** mechanism is first established to extract the initial English key terms from the webpage snippets obtained by using the Chinese OOV term as a query for the search engine. The English fragments segmented by the non-English characters in each snippet are selected. Given the following snippet, "*Naruto wallpapers*", "*Naruto*", "*Two destinys two different fates*" and "*Recognize my existence*" are chosen as the initial key terms.

火影忍者壁纸(Naruto wallpapers)
您的位置：首页 > 火影忍者> 火影忍者壁纸> Naruto ... 火影忍者壁纸： Two destinys two different fates ... 火影忍者壁纸： Recognize my existence ...
www.manmankan.com/wallpaper/1/10/ - 12k - 网页快照 - 类似网页

Obviously, there are a lot of noises among the initial key terms. Therefore, some noise patterns are regarded as **Heuristic Filtering Rules (HFR)** and utilized to remove the noisy strings. (1) If an initial key term appears in the stoplist, then it is removed as a noisy string. The stoplist contains the stopwords with high frequency in common use, which are usually irrelevant with the original OOV term, such as "*Translate this page*" and "*Retrieved from Wikipedia*". (2) If an initial key term begins or ends with a preposition or conjunction, then it is removed as a noisy string. (3) If an initial key term satisfies some filtering patterns, then it is removed as a noisy string. Such patterns are used to select some frequent and obviously incorrect key terms. For example, an initial key term for the OOV term "非洲统一组织 [*Organization of African Unity*]" is "*Fei*1 *zhou*1 *Tong*3 *yi*1 *Zu*3 *zhi*1", which is a unreasonable form composed of both letters and numbers. (4) If multiple initial key terms are same by ignoring the case sensitivity, then the form with the highest frequency is reserved and the others are removed as the noisy strings. For example, for the OOV term "费利克斯[*Felix*]", all the related information for three initial key terms, "*Felix*", "*FELIX*" and "*felix*", must be considered in the subsequent feature selection and computation. (5) For initial key terms with a single word corresponding to the same original OOV term, if a term is a prefix/suffix substring of the other terms, then it is removed as a noisy string.

In the key terms obtained by HFR-based filtering, there are still some redundant substrings, thus the optimization based on **Information Entropy** is proposed to further filter such noises. For a key term x, its entropy is expressed as:

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i) \quad (1)$$

where $p(x_i)$ denotes the frequency of x in the i^{th} snippet, and computed as n_i/n, n_i is the occurrence times of x in the i^{th} snippet and n is the total occurrence times of x in the whole snippet set; N is the total snippet number.

Information Entropy can not only represent the amount of information content for key terms, but also the distribution similarity between two key terms in the snippet set. Given two key terms kt_1 and kt_2, kt_1 is a substring of kt_2. If $\lambda H(kt_1)<H(kt_2)$ (the setting for λ is shown in Section 6.2), then $kt1$ is removed as a noisy string. However, if only using Information Entropy to filter substrings, the relations between an OOV term and its key terms cannot be considered. For key terms with low frequency, they often co-occur with some noisy strings. For example, for the OOV term "萨马兰奇 [*Samaranch*]", its correct translation "*Samaranch*" always occurs in the key term "*Juan Antonio Samaranch*". If only determined by using Information Entropy, "*Samaranch*" will be removed. Thus the special feature *P&S_IF* (defined in Section 4), which describes the phonetic and sense relations between an OOV term and its translation candidates, is added to solve this problem. If ($\lambda H(kt_1)<H(kt_2)$) && ($P\&S\_IF(OOVTerm, kt_1)<P\&S\_IF(OOVTerm, kt_2)$), then kt_1 is deleted.

3 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 a very effective way in the phonetic feature expression. *PV* means that for measuring the edit distance similarity between the syllabification sequences of an OOV term and its candidates, the corresponding 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 and its translation candidate 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.

(3#) **Length Ratio of OOV Term and Its Translation Candidate** (*LR*) – Aims to explore the composition possibility that the translation candidate can be regarded as the final correct 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 "非典型肺炎[*SARS*]" and its candidate "*SARS*" is 3/4=0.75.

(4#) **Phonetic and Sense 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. 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} "斯堪的纳维亚半岛[*Scandinavian Peninsula*]" and its T_{OOV} "*Scandinavian Peninsula*" , the non-transliteration words "半岛[*peninsula*]" and "*Peninsula*" are matched, then *LScore*(S_{OOV}, T_{OOV})=1; the *PV* value between the remaining strings "斯堪的纳维亚[*Scandinavian*]" and "*Scandinavian*" is 0.928, so the final *P&S_IF* value is 1.928/2=0.964.

(5#) **Un-Covered Ratio** (*UCR*) – Aims to explore the ratio of the overlap between an OOV term and the translations of its candidates acquired from Chinese Basic Dictionary (Yang et al. 2009b). It is set up for multi-word OOV terms.

$$UCR(S_{OOV}, T_{OOV}) = 1 - \frac{Len(unTrans)}{Len(S_{OOV})} \quad (4)$$

where *unTrans* is the part in S_{OOV} uncovered by the translation of T_{OOV}. For example, given S_{OOV} "苏伊士运河[*Suez Canal*]" and its T_{OOV} "*Suez Canal*", the part in T_{OOV} which can be translated by Basic Dictionary is "*Canal*" and its translation is "运河[*canal*]". Thus the *unTrans* part in S_{OOV} is "苏伊士[*Suez*]", then the final *UCR* value is 1-3/5=0.4.

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 often occur with the same or similar forms in the latter parts. The contextual information from the same and other Web documents may be beneficial to determine the final translation. To utilize global information, GFs are built 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 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#) **Global Statistical Feature** (*G_SF*) – Aims to explore the statistical measure for the strength of the interdependence between an OOV term and its translation candidates to judge the possibility of a translation candidate being taken as the final correct translation [7].

Chi-Square (χ^2) **Feature Value** (*CV*) – Aims to evaluate the semantic similarity between S_{OOV} and T_{OOV} by their occurrence in Web documents.

$$CV_{\chi^2}(S_{OOV}, T_{OOV}) = \frac{N \times (a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)} \tag{5}$$

where a is the number of snippets with 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 $N=a+b+c+d$.

Information Gain (*IG*) – Aims to compute the probability that T_{OOV} appears in the snippets with S_{OOV}. The larger *IG* shows that T_{OOV} is a more possible translation for S_{OOV}.

$$IG(S_{OOV}, T_{OOV}) = a \times \log \frac{a}{(a+b) \times (a+c)} + b \times \log \frac{b}{(a+b) \times (b+d)} \tag{6}$$

Correlation Coefficient (*CC*) – Aims to measure the linear association degree between S_{OOV} and T_{OOV}. It's a variant of *CV*. The larger *CC* value indicates that the relation between S_{OOV} and T_{OOV} is more correlative, and $CC^2=\chi^2$.

$$CC(S_{OOV}, T_{OOV}) = \frac{\sqrt{N} \times (a \times d - b \times c)}{\sqrt{(a+b) \times (a+c) \times (b+d) \times (c+d)}} \tag{7}$$

Relevance Score (*RS*) – Aims to measure the direct relevance between S_{OOV} and T_{OOV}. It's computed as the ratio between the occurrence probability of T_{OOV} in the snippets with S_{OOV} and that of T_{OOV} in the snippets without S_{OOV}. The larger *RS* indicates that S_{OOV} and T_{OOV} are more relevant.

$$RS(S_{OOV}, T_{OOV}) = \log \frac{\frac{a}{a+b} + m}{\frac{c}{c+d} + m} \tag{8}$$

where m is used to smooth the *RS* and usually set as 1.

Odds Ratio (*OR*) – Aims to measure the indirect relevance between S_{OOV} and T_{OOV}. The distribution of features on relevant candidates is different from that on irrelevant candidates. The larger *OR* indicates that S_{OOV} and T_{OOV} are more relevant.

$$OR(S_{OOV}, T_{OOV}) = \frac{\frac{a}{a+b} \times \left(1 - \frac{c}{c+d}\right)}{\left(1 - \frac{a}{a+b}\right) \times \frac{c}{c+d}} \tag{9}$$

GSS Coefficient (*GSS*) – Aims to measure the relevance between S_{OOV} and T_{OOV}. It is another simplified variant of *CV*. The larger GSS also represents the stronger relevance.

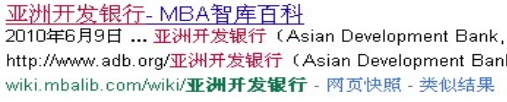
$$GSS(S_{OOV}, T_{OOV}) = a \times d - b \times c \tag{10}$$

(3#) **Pointwise Mutual Information** (*PMI*) – Aims to evaluate the co-occurrence relation between an OOV term and its candidates. If both appear with the higher co-occurrence frequency in the same snippet, they are more relevant.

$$PMI(S_{OOV}, T_{OOV}) = \frac{N \times a}{(a+b) \times (a+c)} \tag{11}$$

(4#) **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 ($p3$). For example, in the following snippet, S_{OOV} is "亚洲开发银行[*Asian Development Bank*, *ADB*]" and T_{OOV} is "*Asian Development Bank*".



$p1_{SOOV}$=0, $p2_{SOOV}$=29, $p3_{SOOV}$=159; $p1_{TOOV}$=36, $p2_{TOOV}$=101, $p3_{TOOV}$=101

The position is indexed from 0. 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:

$$Dist(S_{OOV}, T_{OOV}) = \begin{cases} pi_{S_{OOV}} - pj_{T_{OOV}} - Len(T_{OOV}), & pi_{S_{OOV}} > pj_{T_{OOV}} \\ pj_{T_{OOV}} - pi_{S_{OOV}} - Len(S_{OOV}), & pi_{S_{OOV}} < pj_{T_{OOV}} \end{cases} \quad (12)$$

Given the example above, $Dist=p2_{SOOV}-p1_{TOOV}-6=36-29-6=1$, S_{OOV} and T_{OOV} are a left bracket '(' apart. Thus the average distance *CO-Dist* in the snippet set is:

$$CO\_Dist(S_{OOV}, T_{OOV}) = AVG\_Dist(S_{OOV}, T_{OOV}) = \frac{Sum(Dist)}{CO\_Freq(S_{OOV}, T_{OOV})} \quad (13)$$

where *Sum*( ) is the sum of *Dist* in each snippet.

(5#) **Rank Value** (*RV*) – Aims to consider the rank for translation candidates in the Web document set. It includes six 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.

$$A\_Rank(T_{OOV}) = \frac{Sum(Rank)}{DF_{T_{OOV}}(T_{OOV})} \quad (14)$$

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

$$R\_Rank(T_{OOV}) = \beta \times \frac{|T_{OOV}|}{MAX\_WL} + (1-\beta) \times \frac{TF_{T_{OOV}}(T_{OOV})}{Freq_{S_{OOV}}(S_{OOV})} \quad (15)$$

where β is set as 0.25 empirically, $|T_{OOV}|$ is the length of T_{OOV}, and *MAX_WL* denotes the maximum length of candidates. ***DF_Rank*** (*D_Rank*) is similar to *S_Rank*, and $D\_Rank(T_{OOV})=DF_{TOOV}(T_{OOV})*Len(T_{OOV})$. ***TF_Rank*** is computed as $TF\_Rank(T_{OOV})=TF_{TOOV}(T_{OOV})$, which aims at investigating the impact of the frequency of T_{OOV}.

(6#) **Similarity of Context Vector** (*SCV*) – Aims to evaluate the distribution similarity between an OOV term and its candidates in the snippet set. The OOV term S_{OOV} and its candidate T_{OOV} are first represented as two context vectors, $CV_{SOOV}=(ts_1, \ldots, ts_i, \ldots, ts_N)$ and $CV_{TOOV}=(tt_1, \ldots, tt_i, \ldots, tt_N)$, ts_i and tt_i denote the number of S_{OOV}s and T_{OOV}s in the i^{th} snippet respectively. Thus the *SCV* can be computed as:

$$SCV(S_{OOV}, T_{OOV}) = \cos(CV_{S_{OOV}}, CV_{T_{OOV}}) = \frac{\sum_{i=1}^{N}(ts_i \times tt_i)}{\sqrt{\sum_{i=1}^{N}(ts_i)^2} \times \sqrt{\sum_{i=1}^{N}(tt_i)^2}} \quad (16)$$

Boolean Feature (BF) is a binary feature and equivalent to a heuristic rule designed for the particular relations between an OOV term and its translation candidates. BFs are used to explore the different occurrence forms with higher possibility for the candidates in Web documents. (1#) **Position Distance with OOV Term** (*PD_S*$_{OOV}$) – If T_{OOV} occurs close to S_{OOV} (within 10 characters), this feature is set as 1. (2#) **Neighbor Relation with OOV Term** (*NR_S*$_{OOV}$) – If T_{OOV} occurs prior or next to S_{OOV}, this feature is set as 1. (3#) **Bracket Neighbor Relation 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})", this feature is set as 1. (4#) **Special Mark Word** (*SMW*) – Within a certain co-occurrence distance (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 …*]" or "(或/又)称为[(*or*/*also*) *be called as …*]", or their English translation terms and so on, this feature is set as 1. (5#) **Capitalized First Letter** (*CFL*) – If T_{OOV} begins with a capitalized letter, this feature is set as 1.

4 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 all 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.

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

$$f\left(S_{OOV}, T_{OOV\,i}, w\right) > _{S_{OOV}} f\left(S_{OOV}, T_{OOV\,j}, w\right) \tag{17}$$

where w is a parameter and represented as a vector $\{w_1, \ldots, w_i, \ldots, w_n\}$. This function can also be expressed as:

$$f(S_{OOV}, T_{OOV}, w) = \sum_{k=1}^{p} w_k LF_k(S_{OOV}, T_{OOV}) + \sum_{l=p+1}^{q} w_l GF_l(S_{OOV}, T_{OOV}) + \sum_{m=q+1}^{n} w_m BF_m(S_{OOV}, T_{OOV}) \tag{18}$$

where $LF_k(\ ,\ )$, $GF_l(\ ,\ )$ and $BF_m(\ ,\ )$ are the local, global and Boolean feature representation respectively. These three kinds of feature representation can be incorporated as a whole and represented as a feature function family with the multi-dimensional feature vector in Formula (19).

$$f\left(S_{OOV}, T_{OOV}, w\right) = w \cdot h\left(S_{OOV}, T_{OOV}\right) \tag{19}$$

Thus the relevance for each feature vector x (translation candidate) containing a group of features can be evaluated.

5 Experiment and Analysis

4,170 NEs are selected from the Chinese NER corpus in SIGHAN2005. The test set contains 310 Person Names (PRNs), 324 Location Names (LCNs) and 252 Organization Names (OGNs), and the remaining is taken as the training set. 300 Chinese new terms chosen randomly from 9 categories (movie name, book title, brand name, terminology, idiom, rare animal name and NE), are used to investigate the generalization ability of our model. *Top-N-Inclusion-Rate* is defined as the percentage of the OOV terms whose correct translations could be found in the first *N* translation candidates.

To verify the effectiveness for multiple feature fusion, the test on the feature combination for our model is implemented. As shown in Table 1, the highest *Top*-1-*Inclusion-Rate* of 88.8889% can be acquired by using all the features. It can be seen from Table 1 that the most important features are *P&S_IF*, NR_S_{OOV}, BNR_S_{OOV} and *UCR*. As for the frequency feature, its contribution is limited, because many candidates with higher *P&S_IF* values are the terms with low frequency. However, when training based on only the features that are beneficial to the whole performance, the best translation accuracy is 85.8024%, which is worse than that by combining all the features. Multiple feature fusion can indeed improve the translation accuracy.

Feature				*Top*-1-*Inclusion Rate*	Reduction
All Features				88.8889%	—
Numerical Feature	***Local Numerical Feature***	***-Len***		88.8889%	0.0%
		-PV		84.8765%	-4.01234%
		-LR		88.8889%	0.0%
		-P&S_IF		81.1728%	-7.7160%
		-UCR		84.2592%	-4.6296%
	Global Numerical Feature	***Global Frequency***	$-TF_{TOOV}$	88.8889%	0.0%
			$-DF_{TOOV}$	90.1234%	+1.2345%
			-CO_Freq	89.1975%	+0.3086%
		-CV		88.8889%	0.0%
		-IG		84.5679%	-4.3210%
		-CC		88.8889%	0.0%
		-RS		85.1852%	-3.7037%
		-OR		89.8148%	+0.9259%
		-GSS		88.8889%	0.0%
		-PMI		89.8148%	+0.9259%
		-CO_Dist		87.0370%	-1.8518%
		RV	***-T_Rank***	88.2716%	-0.6172%
			-A_Rank	89.8148%	+0.9259%
		-SCV		89.5062%	+0.6173%
Boolean Feature		$-PD\_S_{OOV}$		88.2716%	-0.6173%
		$-NR\_S_{OOV}$		83.6419%	-5.2469%
		$-BNR\_S_{OOV}$		83.9506%	-4.9383%
		-SMW		88.8889%	0.0%
		-CFL		89.1975%	+0.3086%

Table 1. Results for feature combination.

Yang et al. [23] is very similar to our approach, we accomplished this method on the same data set to make a contrast, as shown in Table 2. It can be concluded that the ranking based on the supervised learning outperforms the existing conventional strategies, Ranking SVM is better than SVM for ranking, and our approach is superior to Yang et al.'s. Meanwhile, the best performance is obtained for PRNs. It shows that our model is sensitive to the category and the popularity of OOV term.

Method	Ranking Pattern	Category	*Top*-1	*Top*-2	*Top*-3
Our Model	***based on SVM* (*Multiple Features*)**	***PRN***	88.70%	97.09%	99.35%
		LCN	76.23%	93.82%	96.91%
		OGN	76.58%	92.06%	96.42%
		All	**80.69%**	**94.46%**	**97.62%**
	***based on Ranking SVM* (*Multiple Features*)**	***PRN***	92.58%	97.74%	99.03%
		LCN	87.34%	95.37%	98.14%
		OGN	84.52%	95.23%	97.22%
		All	**88.89%**	**96.16%**	**98.19%**
Yang *et al.* [23]	***based on SVM* (TF_{TOOV}+*LR*+*UCR*+*CFL*)**	***OGN* (Only)**	53.96%	76.98%	88.49%
	***based on Ranking SVM* (TF_{TOOV}+ *LR*+*UCR*+*CFL*)**	***OGN* (Only)**	62.69%	83.33%	88.49%

Table 2. Performance comparison results.

Another test for the other kinds of Chinese OOV term is performed on the selected new terms and the consistent results can be observed in Table 3.

Top-N-Inclusion-Rate	*Top*-1	*Top*-3	*Top*-5	*Top*-7	*Top*-9
Chinese OOV New Terms	74.66%	90.33%	94.33%	95.00%	96.00%

Table 3. Results for Chinese OOV new terms.

Four CLIR runs are carried out on the Chinese topic set and English corpus from TREC-9. (1) *C-E_LongCLIR*1 – using Long Query (LQ, terms in both title and description fields) and the Dictionary-Based Translation (DBT); (2) *C-E_LongCLIR*2 – using LQ, DBT and our model; (3) *C-E_ShortCLIR*1 – using Short Query (SQ, only terms in the title field) and DBT; (4) *C-E_ShortCLIR*2 – using SQ, DBT and our model. The Precision-Recall curves and Median Average Precision (MAP) are shown in Fig. 1. It can be seen from Fig. 1 that the best run is *C-E_LongCLIR*2, and its results exceed those of *C-E_LongCLIR*1. By adopting both query translation based on bilingual dictionary and OOV term translation, Chinese-English CLIR for long query has gained the significant retrieval performance improvement. The same conclusion can be obtained for the other two runs *C-E_ShortCLIR*1 and *C-E_ShortCLIR*2.

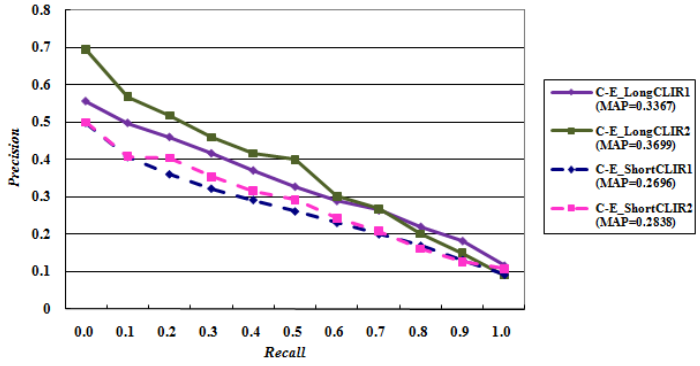


Fig. 1. Results for Chinese-English CLIR combining our model.

Through analyzing the results, 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 "经济法制化", the mining result based on Google in China is "*to manage economic affairs according to 1aw*", which is more reasonable than "*Economic law*" acquired by Bing. (2) S**ome terms are idioms, conventional and political terminologies with Chinese characteristics, and cannot be translated literally.** For example, "党群关系[*party masses relationship*]" should be translated into "*party masses relationship*", rather than "*ties between the party*" 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 an obvious difference among the translation performance for PRN, LCN and OGN. (4) **There are some particular and inherent noises in the extracted translation candidates.** For example, a candidate for the Chinese OOV term "广东人民出版社[*Guangdong People's Publishing House*]" is "*Guangdong ren min chu ban she*". (5) **Word Sense Disambiguation (WSD) should be added to improve the translation performance.** Although most of OOV terms have a unique sense definition, there are still a few OOV terms with sense ambiguity, e.g., "东北大学[*Northeastern University* or *Tohoku University*]".

6 Conclusions

Traditional OOV term translation methods concern two aspects, that is, transliteration and sense translation. However, more and more Chinese OOV terms cannot be measured by phonetic or meaning information separately. Our proposed model improves the acquirement ability for Chinese-English 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. Our future research will focus on applying the key techniques on statistical machine learning, alignment of sentence and phoneme, and WSD into Chinese-English OOV term translation.

Acknowledgments. This work is supported by National Science & Technology Pillar Program of China (No. 2012BAH59F04), National Natural Science Fund of China (No. 61170095; No. 71171126), and Shanghai Municipal R&D Foundation (No. 12dz1500203, 12511505300). Cheng Jin is the corresponding author.

References

1. Al-Onaizan Y., and Knight K. 2002. Translating Named Entities using Monolingual and Bilingual Resources. In *Proceedings of ACL 2002*, 400-408.
2. Cao Y.B., Xu J., Liu T.Y., Li H., Huang Y.L., and Hon H.W. 2006. Adapting Ranking-SVM to Document Retrieval. In *Proceedings of SIGIR 2006*, 186-193.
3. Chen C., and Chen H.H. 2006. A High-Accurate Chinese-English NE Backward Translation System Combining Both Lexical Information and Web Statistics. In *Proceedings of COLING-ACL 2006*, 81-88.
4. Fang G.L., Yu H., and Nishino F. 2006. Chinese-English Term Translation Mining based on Semantic Prediction. In *Proceedings of COLING-ACL 2006*, 199-206.

5. Ge Y.D., Hong Y., Yao J.M., and Zhu Q.M. 2010. Improving Web-Based OOV Translation Mining for Query Translation. In *Proceedings of AIRS 2010*, LNCS 6458, 576-587.
6. Hu R., Chen W., Bai P., Lu Y., Chen Z., and Yang Q. 2008. Web Query Translation via Web Log Mining. In *Proceedings of SIGIR 2008*, 749-750.
7. Huang S., Chen Z., Yu, Y., and Ma W.Y. 2006. Multitype Features Coselection for Web Document Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):448-459.
8. Jiang L., Zhou M., Chien L.F., and Niu C. 2007. Named Entity Translation with Web Mining and Transliteration. In *Proceedings of IJCAI 2007*, 1629-1634.
9. Joachimes T. 2002. Optimizing Search Engines using Click through Data. In *Proceedings of SIGKDD 2002*, 133-142.
10. Lee C.J., Chang J.S., and Jang J.R. 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.
11. Lu W.H., and Chien L.F. 2002. Translation of Web Queries using Anchor Text Mining. *ACM Transactions on Asian Language Information Processing*, 1(2):159-172.
12. Lu W.H., and Chien L.F.. 2004. Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach. *ACM Transactions on Information Systems*, 22(2):242-269.
13. Ren F.L., Zhu M.H., Wang H.Z., and Zhu J.B. 2009. Chinese-English Organization Name Translation Based on Correlative Expansion. In *Proceedings of the 2009 Named Entities Workshop*, *ACL-IJCNLP 2009*, 143-151.
14. Shao L., and Ng H.T. 2004. Mining New Word Translations from Comparable Corpora. In *Proceedings of COLING 2004*, 618-624.
15. Shi L. 2010. Mining OOV Translations from Mixed-Language Web Pages for Cross Language Information Retrieval. In *Proceedings of ECIR 2010*, LNCS 5993, 471-482.
16. Sproat R., Tao T., and Zhai C.X. 2006. Named Entity Transliteration with Comparable Corpora. In *Proceedings of COLING-ACL 2006*, 73-80.
17. Virga P., and Khudanpur S. 2003. Transliteration of Proper Names in Cross-Language Applications. In *Proceedings of SIGIR 2003*, 365-366.
18. Wang J.H., Teng J.W., Cheng P.J., Lu W.H., and Chien L.F. 2004. Translating Unknown Cross-Lingual Queries in Digital Libraries using a Web-based Approach. In *Proceedings of JCDL 2004*, 108-116.
19. Wu J.C., and Chang J.S. 2007. Learning to Find English to Chinese Transliterations on the Web. In *Proceedings of EMNLP-CoNLL 2007*, 996-1004.
20. Xu J., Cao Y.B., Li H., and Zhao M. 2005. Ranking Definitions with Supervised Learning Methods. In *Proceedings of WWW 2005*, 811-819.
21. Yang F., Zhao J., Zou B., and Liu K. 2008. Chinese-English Backward Transliteration Assisted with Mining Monolingual Web Pages. In *Proceedings of ACL 2008*, 541-549.
22. Yang F., Zhao J., and Liu K. 2009a. A Chinese-English Organization Name Translation System Using Heuristic Web Mining and Asymmetric Alignment. In *Proceedings of ACL-AFNLP 2009*, 387-395.
23. Yang M., Shi Z.. Li S., Zhao T., and Qi H. 2009b. Ranking vs. Classification: a Case Study in Mining Organization Name Translation from Snippets. In *Proceedings of IALP 2009*, 308-313.
24. Zhang Y., Huang F., and Vogel S. 2005. Mining Translations of OOV Terms from the Web through Cross-Lingual Query Expansion. In *Proceedings of SIGIR 2005*, 669-670.
25. Zhang Y. and Vines P. 2004. Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In *Proceedings of SIGIR 2004*, 162-169.