# Web based English-Chinese OOV term translation using Adaptive rules and Recursive feature selection

Jian Qu, Nguyen Le Minh, Akira Shimazu

School of Information Science, JAIST
Ishikawa, Japan 923-1292
qujian@jaist.ac.jp

**Abstract.** Cross-Language Information Retrieval (CLIR) system uses dictionaries for information retrieval. However, out of vocabulary (OOV) terms cannot be found in dictionaries. Although many researchers in the past have endeavored to solve the OOV term translation problem, but little attention has been paid to hybrid translations "α1-antitrypsin deficiency (α1-抗胰蛋白酶缺乏症)". This paper presents a novel OOV term translation mining approach, which proposes a new adaptive rules system for hybrid translations and a new recursive feature selection method for supervised machine learning. We evaluate the propose method with English-Chinese OOV term translation. Our experiments show that adaptive rules system and recursive feature selection with Bayesian Net can significantly outperform existing supervised models.

**Keywords:** adaptive rules, recursive feature selection.

## 1 Introduction

CLIR system employs multilingual dictionaries for translating queries from one to another. To ensure a perfect CLIR system, it must use a perfect multilingual dictionary. However, a perfect multilingual dictionary would never be existed due to OOV terms. OOV terms are typically new terms that cannot be found in dictionaries, such as personal names, place names, new technical terms and translated words etc. New OOV terms are emerging every day, this makes it difficult for multilingual dictionaries to cover all translations of OOV terms.

Generally speaking, OOV terms can be translated by three methods, they are: human translation with field of specified knowledge; automatic translation from parallel corpus or comparable corpus (Zhou, Truran et al. 2008); and automatic mining from the Internet (Lu, Chien et al. 2004). Most OOV terms have their correspondent human translation nearby on the Internet (Cheng, Teng et al. 2004; Zhang, Huang et al. 2005; Udupa, K et al. 2009). The quality of the translation is incomparable between human translation and automatic translation. However, automatic translation offers the speed and low cost over human translation. While automatic mining takes the advantage of both human and automatic translation. The translations mined from the Internet are usually high quality and require low man power cost. According to our observation, some translations of the technical OOV term are hybrid translations. Hybrid translations use part target language and part source language, for example, a English OOV term: "DiGeorge's syndrome" and its Chinese translation "DiGeorge's 症候群". "DiGeorge's" are source languages and "症候群" are target languages. Solving Hybrid translations is important because existing approaches would retrieve "症候群" as the translation of "DiGeorge's syndrome". If this translation is applied to CLIR, many disease documents unrelated to "DiGeorge's syndrome" will be retrieved, because many Chinese medical terms end with the term "症候群". In this paper, we propose an OOV term translation method that employs the automatic mining, a novel adaptive rules system for hybrid translations, and a novel recursive feature selection using supervised machine learning by Bayesian net with Adaboost.

## 2 Related Research Work

Many researchers in the past have endeavored to solve the OOV term translation problem (Cheng, Teng et al. 2004; Lu, Chien et al. 2004; Zhang and Vines 2004; Zhang and Vines 2004; Zhang, Vines et al. 2005; Yuejie, Yang et al. 2009). However, new OOV terms are emerging every day, a perfect method that is able to handle all OOV terms is yet to be discovered. Existing methods for name type OOV terms usually cannot handle technical type OOV terms (Zhang and Vines 2004). Zhang and Vines extract 30 Chinese characters before and after the OOV term when source OOV term is found, then they use lengths and frequencies of translation candidates to select the correct translation. Most researchers approach the OOV term problems with brute force translation candidate extraction (Cheng, Teng et al. 2004; Zhang and Vines 2004), this approach may generate many noises, and it is difficult to handle technical type OOV terms. Especially for hybrid type translations which use part target language and part source language "DiGeorge's syndrome (DiGeorge's 症候群)", "α1-antitrypsin deficiency (α1-抗胰蛋白酶缺乏症)". Recently, many researchers started to translate the OOV terms using supervised machine learning (Tiffin, Kelso et al. 2005). Never the less feature selection has always been a problem for supervised machine learning. Traditional feature selections may be time and memory consuming to get high performance. Many supervised machine learning approaches for OOV term translation use support vector machine (SVM) as the classifier for translation selection (Yuejie, Yang et al. 2009), Yuejie, Yang *et al.* use PAT-tree and local maximum algorithm to extract the Chinese translations, then use SVM to select the correct translations. However, SVM works well with OOV terms that have similar properties. Technical OOV terms have very different properties, some can be found in over 100 webpage on the Internet, while others can be found in as little as 3 webpages on the Internet. We propose an OOV term translation method with a novel adaptive rules system for translation candidate extraction to reduce the noise and to handle hybrid translations. We also propose a new recursive feature selection method to gain the advantage of better feature set and high speed with low memory usage. We employee Bayesian Net which can handle diversity of OOV terms together with Meta level-Adaboost that helps with over fitting.

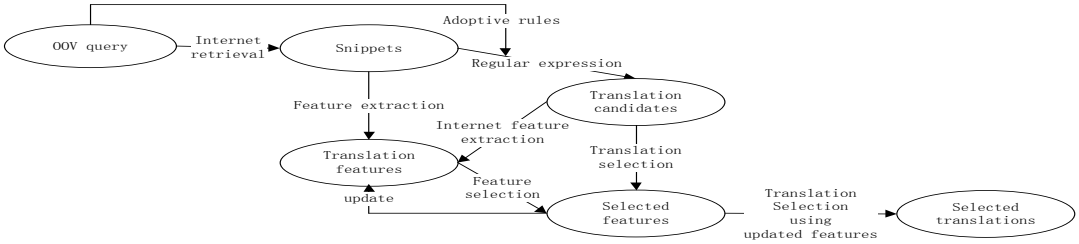


**Figure 1:** Flow chart of our proposed approach

## 3 The Proposed Method

Our approach is developed into four parts, they are: firstly, the Internet retrieval; secondly, the translation candidate extraction; thirdly, the feature extraction and finally the translation selection. A flow chart of our approach is shown in Figure 1.

### 3.1 Internet Retrieval

We feed the complete string of OOV terms to Yahoo API(Yahoo 2009) and limit the result in Chinese language. An example of snippet containing OOV term and its translation is shown in Figure 2.

α1-抗胰蛋白酶缺乏症|症状|治疗 (Title)

2009 年 1 月 10 日 ... α1-抗胰蛋白酶缺乏症(α1-antitrypsin deficiency)是以婴儿期出现胆汁 ... 的糖蛋白，在化学组成上与正常α1-AT 的区别是缺乏唾液酸基和糖基。 ... (Summary)

www.yongyao.net/jbhtml/α1-kydbmqfz.htm (URL)

**Figure 2:** An example of web retrieved snippets of "α1-antitrypsin deficiency"

Unlike many existing approaches, HTML tags are kept to assist translation candidate extraction.

## 3.2 Translation Candidate Extraction

The translation candidate extraction is based on the idea that most OOV terms have their correspondent human translation nearby (Cheng, Teng et al. 2004; Zhang, Huang et al. 2005; Udupa, K et al. 2009). According to our observation, we found out some hybrid translations of the OOV terms may not only use the target language alphabets or characters, but also use some alphabets, characters or symbols from the source language. We propose to include the alphabets, characters or symbols of the source language by using an adaptive rules system.

The adaptive rules system uses a set of predefined regular expression matching rules as the base rules. The base rules are modified by each OOV term to form the adapted regular expression matching rules for translation candidate extraction. The base regular expression matching rules are shown in Table 1.

**Table 1:** Regular expression matching rules

| # | Regular expression matching rules |
|---|---|
| 1 | Chinese characters |
| 2 | Chinese characters/Other characters |
| 3 | Chinese characters/Other characters/Chinese characters |
| 4 | Chinese characters/Other characters/Chinese characters/Other characters |
| 5 | Other characters/Chinese characters |
| 6 | Other characters/Chinese characters/Other characters |
| 7 | Other characters/Chinese characters/Other characters/Chinese characters |
| *Note:* Chinese characters = all Chinese characters<br>Other characters = all non-Chinese language alphabets, characters, symbols. | |

The adaptive rules system is developed as follows. Let *Sn* be the snippets retrieved from the Internet, OOV be the source OOV terms, *A* be any alphabets, characters or symbols, *Ac* be any Chinese characters, Re be regular expression matching rules, *Ar* be adopted matching rules, and *Tc* be Chinese translation candidates. For each OOV term, if it is found in the snippets, we add the substring of the OOV term to the regular expression matching rules to create the adopted matching rules. Then we scan for the nearest Chinese character in front of or after the OOV terms. Once we find the Chinese character, we try to match the string around the Chinese character with the adopted matching rules. If there are one or more rules that match the string, we extract the matched parts of the string as the Chinese translation candidates. The detailed algorithm of this method is explained below.

**Algorithm Translation candidate extraction**

***Input:*** *Snippets retrieved from the Internet Sn, OOV terms OOV, Any alphabets, characters or symbols A, Any Chinese characters Ac, Regular expression matching rules Re,*

***Output:*** *Adopted matching rules Ar, Chinese Translation candidates Tc,*

```
For each OOV found in Sn do
Ar = Re + Substring OOV,
  If (Ac found in front or behind OOV)then
  Continue;
   If(Ar found near Ac+A)then
      Matching Ar with Ac+A;
                    Tc = Ac+A;
     End
  End
End
```

An example of translation candidate extraction is shown in Table 2. As can be seen from this example, the OOV term is α 1-antitrypsin deficiency, we created the adopted matching rules using the substring of the OOV and the base regular expression matching rules in Table 1. The correct translation is extracted as α 1-抗胰蛋白酶缺乏症.

**Table 2:** Example of translation candidate extraction

| | | |
|---|---|---|
| OOV | α 1-antitrypsin deficiency | |
| Snippet | 2009 年 1 月 10 日 ... α 1-抗胰蛋白酶缺乏症( α 1-antitrypsin deficiency)是以婴儿期出现胆汁 ... 的糖蛋白，在化学组成上与正常 α 1-AT 的区别是缺乏唾液酸基和糖基。 | |
| Adopted matching rules | 1 | Chinese characters |
| | 2 | Chinese characters/ α /1/-/antitrypsin/deficiency/ |
| | 3 | Chinese characters/ α /1/-/antitrypsin/deficiency/Chinese characters |
| | 4 | Chinese characters/ α /1/-/antitrypsin/deficiency/Chinese characters/ α /1/-/antitrypsin/deficiency |
| | 5 | α /1/-/antitrypsin/deficiency/Chinese characters |
| | 6 | α /1/-/antitrypsin/deficiency/Chinese characters/ α /1/-/antitrypsin/deficiency/ |
| | 7 | α /1/-/antitrypsin/deficiency/Chinese characters/ α /1/-/antitrypsin/deficiency/Chinese characters |
| Matched translations | α 1-抗胰蛋白酶缺乏症<br>是以婴儿期出现胆汁 | |

## 3.3 Feature Extraction

In this subsection, we extracted totally 24 different features from translation candidates. These features include: average distances, co-occurrence distance, term frequencies, symmetric conditional probability (SCP), modified association measures, chi-square, lengths of OOV and translation, and length similarity. We describe the details of these features as follows.

*1)* Distances between OOV and translation

The closer a translation candidate to its source OOV term the more likely that translation is correct (Cheng, Teng et al. 2004). Some translations occur in front and after the OOV term, but some translations only occur in front or after the OOV term. To present the actual locations between the OOV and translations, we need to consider the average distance $Dist(c_i e_i)$, average front distance $Dist(c_i, e_i)$ and the average back distance $Dist(e_i, c_i)$.

*2)* Co-occurrence distance

Co-occurrence distance ($CDist$) is the sum of average distance between OOV and translation candidate over the co-occur frequency between OOV and translation candidate. It is computed as follows.

$$CDist = \frac{sum(Dist(c_i e_i))}{tf(c_i e_i)} \tag{1}$$

A modification of this feature ($CwDist$) was proposed by (Yuejie, Yang et al. 2009), they use the web retrieved page count instead of the $tf(c_i e_i)$. It is computed as follows.

$$CwDist = \frac{sum(Dist(c_i e_i))}{S(c_i)} \tag{2}$$

Equation (1) shows the calculation of $CDist$, where $tf(c_i e_i)$ is the co-occur frequency between OOV and translation candidate. Equation (2) shows the calculation of $CwDist$, where $S(c_i)$ is the web retrieved page count of translation candidate.

*3)* Term frequencies

Term Frequencies are very important statistical features for OOV term translation, the more a translation co-occurs with an OOV term the more likely to be the correct translation. We collect

the term frequencies of the translation candidates $tf(c_i)$, OOV $tf(e_i)$, and the co-occur frequencies of translations and OOV $tf(c_i e_i)$. Furthermore, to cope with the average front distance and average back distances, we also collect the front frequency $tf(c_i, e_i)$ and back frequency $tf(e_i, c_i)$ for the translation candidates.

*4)* Symmetrical Conditional Probability

Symmetrical Conditional Probability ($SCP$) (Silva, Jos et al. 1999; Cheng, Teng et al. 2004; Lu, Xu et al. 2007) checks each alphabet, character and substring in the possible translation. By calculating the frequency of each substring in the corpus and compare them to the frequency of the translation, it results higher if the substrings of the translation occur less often in the corpus than they occur only within the translation itself. If a translation has higher SCP value, the translation is more likely to be a word phrase and less likely to be a sentence.

$$SCP(c_1 \dots c_n) = \frac{(n-1) f(c_1 \dots c_n)^2}{\sum_{i=1}^{i=n} f(c_1 \dots c_i) f(c_{i+1} \dots c_n)} \tag{3}$$

Equation (3) shows the calculation of SCP, where $(c_1 \dots c_n)$ is any possible translation candidates, $n$ is the number of characters in this translation candidate, $f(c_1 \dots c_n)$ is the frequency of the translation candidate, and $f(c_1 \dots c_i)$ or $f(c_{i+1} \dots c_n)$ is the frequency of any substring of the translation candidate.

*5)* Modified Association Measures

We propose the modified association measures, which do not require the total number of pages in the Internet. They take the webpage count of OOVs $S(e_i)$, translations $S(c_i)$, the webpage count of OOVs co-occur with translations $S(e_i \wedge c_i)$ and the webpage count of OOVs occur without translations $S(e_i \wedge \neg c_i)$ from the Internet. These features utilize the search engines to remove some possible wrong translation candidates, because search engines use some predefined segmentation tools to eliminate the meaningless Chinese strings.

*a)* ***Support***

Support is an undirected measure that finds the ratio when $e_i$ and $c_i$ occur together in a same webpage, and it is computed as follow.

$$Supp(e_i \rightarrow c_i) = S(e_i \wedge c_i) \tag{4}$$

*b)* ***Confidence***

Confidence is a directed measure for the ratio that $e_i$ occurs when $c_i$ is occurred, and it is computed as follow.

$$Conf(e_i \rightarrow c_i) = \frac{S(e_i \wedge c_i)}{S(e_i)} \tag{5}$$

*c)* ***Lift or Interestingness***

Lift or interestingness takes the correlation between $e_i$ and $c_i$, it tests on two hypotheses, first $e_i$ is independent of $c_i$, so they do not co-occur in the same webpage. Second $e_i$ and $c_i$ are depended and correlated in the same webpage, and it is computed as follow.

$$lift(e_i \rightarrow c_i) = \frac{S(e_i \wedge c_i)}{S(e_i) S(c_i)} \tag{6}$$

*d)* ***Conviction***

Conviction represents the ratio of the expected frequency that $e_i$ occurs without $c_i$, and it is computed as follow.

$$Conv(e_i \rightarrow c_i) = \frac{S(e_i)(\neg c_i)}{S(e_i \wedge \neg c_i)} \tag{7}$$

Equation (4) is the Support of association measure, equation (5) is the Confidence of the association measure, equation (6) is the Lift of the association measure, and equation (7) is the Conviction of the association measure. $e_i$ is the OOV term, $c_i$ is the translation and $S(e_i)$ is the number of pages returned by the search engine when $e_i$ is submitted as a query. $(\neg c_i)$ is assumed to be 1, because translations of OOV terms have a very small portion when compare to the whole Internet.

*6)* Chi-square

Chi-square tests a list of possible translation candidates with their source OOV term. A correlation relationship between the OOV term and its translation candidates can be measured by this method.

$$\boldsymbol{\chi^2}(e_i, c_i) = \frac{N \times (a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)} \tag{8}$$

Equation (8) is the Chi-square tests, where the meaning of each variable is explained as follows

| | | | | | |
|---|---|---|---|---|---|
| $e_i$ | = | source OOV term | b | = | $S(e_i \wedge \neg c_i)$ |
| $c_i$ | = | translation candidates | c | = | $S(c_i \wedge \neg e_i)$ |
| a | = | $S(e_i \wedge c_i)$ | d | = | $S(\neg e_i \neg c_i)$ |
| | | | S() | = | Is a function that takes a query as the input and returns the number of snippets in the corpus containing that query |

*7)* Length of OOV and translation candidates

The translation of OOV should have similar ratio of length, we collect the lengths of translation candidates ($|c_i|$), lengths of OOV terms ($|e_i|$) and the differences between them D($|e_i|, |c_i|$).

*8)* Length similarity

Other than simple differences between the lengths of OOV and the translations, we also employee the length similarity ratio, it is a normalized length difference δ($|e_i|, |c_i|$). and it is computed as follow (Shi 2010).

$$\delta(|e_i|, |c_i|) = \frac{|c_i| - |e_i| \times c}{\sqrt{(|e_i| + 1)\sigma^2}} \tag{9}$$

Equation (9) is the length similarity, where c is a constant indicating the average length ratio between OOV and translations. $\sigma^2$ is the variance of D($|e_i|, |c_i|$).

### 3.4 Translation Candidate Selection

In this section, we explain our candidate selection method. It is developed into two parts, the statistical filter, and the Bayesian net candidate selection with recursive feature selections.

One OOV term can retrieve up to few hundreds of translation candidates, most of them are substrings of the correct translation, and some of them are the longer strings of the correct translation. Two features in our feature set can simply filter some wrong candidates, they are co-occur frequency ($tf(c_i e_i)$) and location distance ($Dist(c_i e_i)$). Both features are very important to the candidate selection, if a Chinese translation co-occurs very often with the source English OOV term and this translation is found very close to the source English OOV term, then this translation may less likely be the wrong translation (noise). Our filter takes the top 70% of the co-occur frequency and the shortest location distance between OOV and the translations. A recall test was performed to evaluate the setting of this filter, the result is shown in Figure 3.

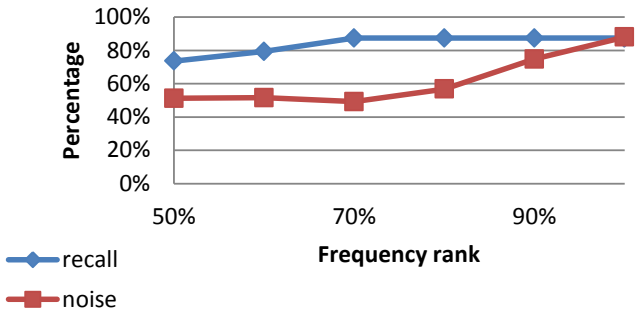


**Figure 3**: Recalls of frequency rank

To handle the diversity of OOV terms, we employed the Bayesian net which can establish an inference reasoning and causality between OOV terms and the translation candidates. We also employ the Meta level-Adaboost to handle the over fitting problems (Rapidminer 2009).

In order to achieve a low memory usage and high performance feature selection method, we developed a recursive feature selection method. This method loops backward elimination by updating the inputting feature sets with the selected features from the backward elimination. This method is developed as follows. Let *OOV* be OOV terms, *Tc* be translation candidates, *F* be inputting feature sets, *N* be number of iterations, *Ct* be Chinese translations, and *SF* be selected feature sets. We use Bayesian net for candidate selection with *N* times of iterations, backward elimination is employed as the base feature selection method. For each iteration, a set of features will be selected by backward elimination, we then update the inputting feature set for next iteration with the selected features from backward elimination. The detailed algorithm of this approach is shown below.

**Algorithm Recursive feature selections**

***Input:*** *OOV terms OOV, Translation candidates Tc, Inputting Feature sets F, Number of Iterations N.*

***Output:*** *Chinese Translations Ct, Selected Feature sets SF,*
*For i =1 to N do*

*F = F';*
*Bayesian net(OOV, Tc) using backward elimination( F),*
*F' = SF;*
*Ct' = Selected Tc;*
*End*
*Ct = Ct';*

## 4 Experiment and Discussion

In this section, we describe our experiments and discuss the results.

### 4.1 Data Collection

We collected English medical terms from Classification of Diseases, Functioning, and Disability (ICD9 2009). The list is called "International Classification of Diseases, Ninth Revision (ICD-9)". It contains totally 2,741 English Medical terms. We randomly selected 10 percent to reduce the time for human evaluation, because for some OOV terms, the correct translation were not published by Taiwan Center of Disease Controls (Taiwan 2009), and some OOV terms may have few correct translations and not all listed in the Taiwan Center of Disease Controls (Taiwan 2009), so we need to find the correct translation from Internet for evaluation. We extracted any English medical terms that can be found on Chinese webpages according to Yahoo API. These English terms retrieved a total of 12,446 Chinese snippets, and extracted a total of 15,042 possible translation candidates. We also manually queried each English medical term to the web and compare to the published resource of Taiwan Center of Disease Controls (Taiwan 2009) to find the correct Chinese translations for comparison with our experimental results.

### 4.2 Machine learning setup and error analysis tool

RapidMiner (Rapidminer 2009) is used for machine learning and error analysis since it details out each results on learning process. We used 10-fold Cross Validation to experiment on the data collection.

### 4.3 Parameters

We used the following setting in our experiments:

Number of Iterations = 2. Estimate = Simple Estimator.

Searching algorithm = Simulated Annealing.

### 4.4 SVM method from Yuejie and Length co-occurrence method from Zhang

To compare our method with existing methods, we employ the SVM translation selection method from Yuejie, Yang *et al*. (2009) using their feature set to compare with our translation selection method. We also employ the length co-occurrence method from Zhang & Vines (2004) to compare with our translation extraction method.

### 4.5 Mining Results and Discussions

We applied our method and length co-occurrence method from Zhang & Vines to the same data collection and the results are presented in Table 3. As can be seen from Table 3, our adaptive rules can extract many translations where Zhang & Vines' method failed to extract, mostly because we considered the existence of the hybrid type OOV translations. Furthermore, to have a fare comparison between translation selection methods, we used our extracted translations to apply with Yuejie's SVM method and our proposed method for translation selection. The Bayesian net with Adaboost outperforms SVM in translation candidate selection with an accuracy of 91.17%. The recursive feature selection improved the performance mainly due to removal of noise features. However, we also observed two iterations of recursive feature selection provided the best results. More number of iterations would diminish the performance. Table 4 shows which features were selected.

**Table 3**: Comparison of our proposed method with existing methods

| Method | Transliterable OOVs | Correct translations mined | Correct translation mined in OOV terms | Iteration 1 | | | Iteration 2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Correct translations selected | wrong translations selected | Correct translations selected in OOV terms | Correct translations selected in OOV terms | wrong translations selected | Correct translations selected in OOV terms |
| Our method | 238 | 246 | 228(95.79%) | 222 | 31 | 213(89.49%) | 220 | 27 | 217(91.17%) |
| Yuejie's method using our translations | | | | 241 | 102 | 161(67.64%) | | | |
| Zhang & Vines' method | 238 | 217 | 159(66.80%) | 143(60.08%) | | | | | |

**Table 4**: Feature selection of our methods

| # | Features | Iteration 1 | Iteration 2 | # | Features | Iteration 1 | Iteration 2 |
|---|---|---|---|---|---|---|---|
| 1 | $Dist(c_i e_i)$ | × | | 11 | $SCP(c_1 \dots c_n)$ | ✓ | × |
| 2 | $Dist(c_i, e_i)$ | ✓ | ✓ | 12 | $Supp(e_i \rightarrow c_i)$ | ✓ | × |
| 3 | $Dist(e_i, c_i)$ | ✓ | ✓ | 13 | $Conf(e_i \rightarrow c_i)$ | ✓ | ✓ |
| 4 | $CDist$ | ✓ | × | 14 | $lift(e_i \rightarrow c_i)$ | ✓ | ✓ |
| 5 | $CwDist$ | ✓ | × | 15 | $Conv(e_i \rightarrow c_i)$ | ✓ | ✓ |
| 6 | $tf(c_i)$ | ✓ | ✓ | 16 | $\chi^2(e_i, c_i)$ | ✓ | ✓ |
| 7 | $tf(e_i)$ | × | | 17 | $\lvert c_i \rvert$ | × | |
| 8 | $tf(c_i e_i)$ | ✓ | ✓ | 18 | $\lvert e_i \rvert$ | × | |
| 9 | $tf(c_i, e_i)$ | ✓ | ✓ | 19 | $\mathrm{D}(\lvert e_i \rvert, \lvert c_i \rvert)$ | ✓ | ✓ |
| 10 | $\mathrm{tf}(e_i, c_i)$ | ✓ | ✓ | 20 | $\delta(\lvert e_i \rvert, \lvert c_i \rvert)$ | ✓ | ✓ |

## 5 Conclusion and Future work

Cross lingual information retrieval system relays heavily on dictionaries. Many OOV terms cannot be found on these dictionaries. To achieve high performance in CLIR systems, OOV terms need to be consistently updated to these dictionaries. Many researches in the past have endeavored to solve the OOV term translation problem, however they yield low performances with technical OOV term translation, especially hybrid type translations. We proposed a method that uses adaptive rules to handle the translation extraction for hybrid type translations. We also proposed a recursive feature selection for Bayesian net with Ada boost for translation selection. We tested our methods with the ICD9 English medical OOV terms. 12.6% of the experimental data was hybrid translations. Our proposed methods resulted a much higher performance than existing methods with a precision of 91.17%. The findings support the idea that improvements of hybrid translation will improve the overall OOV term translation performance.

We plan to investigate more in adaptive rules for translation candidate extraction, because more improvements on the translation candidate extraction will improve the overall OOV translation performance.

## References

Cheng, P.-J., J.-W. Teng, et al. (2004). Translating unknown queries with web corpora for cross-language information retrieval. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Sheffield, United Kingdom, ACM**:** 146-153.

ICD9, C. (2009). "International Classification of Diseases,

Ninth Revision (ICD-9)." from http://www.cdc.gov/nchs/icd/icd9.htm.

Lu, C., Y. Xu, et al. (2007). Translation disambiguation in web-based translation extraction for English-Chinese CLIR. Proceedings of the 2007 ACM symposium on Applied computing. Seoul, Korea, ACM**:** 819-823.

Lu, W.-H., L.-F. Chien, et al. (2004). "Anchor text mining for translation of Web queries: A transitive translation approach." ACM Transactions on Information and System (ACM Trans. Inf. Syst.) **22**(2): 242-269.

Rapidminer (2009). "Rapidminer data mining tool."

Shi, L. (2010). Mining OOV Translations from Mixed-Language Web Pages for Cross Language Information Retrieval. Advances in Information Retrieval. C. Gurrin, Y. He, G. Kazaiet al, Springer Berlin / Heidelberg. **5993:** 471-482.

Silva, J. F. d., S. G. Jos, et al. (1999). Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence, Springer-Verlag**:** 113-132.

Taiwan, C. o. (2009). "Centers for disease control Taiwan." from http://flu.cdc.gov.tw.

Tiffin, N., J. F. Kelso, et al. (2005). "Integration of text and data-mining using ontologies successfully selects disease gene candidates." Nucleic Acids Res **33**: 1544–1552.

Udupa, R., S. K, et al. (2009). "They Are Out There, If You Know Where to Look": Mining Transliterations of OOV Query Terms for Cross-Language Information Retrieval. Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval. Toulouse, France, Springer-Verlag**:** 437-448.

Yahoo. (2009). "Yahoo API." from http://developer.yahoo.com.

Yuejie, Z., W. Yang, et al. (2009). English-Chinese bi-directional OOV translation based on web mining and supervised learning. Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Suntec, Singapore, Association for Computational Linguistics**:** 129-132.

Zhang, Y., F. Huang, et al. (2005). Mining translations of OOV terms from the web through cross-lingual query expansion. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. Salvador, Brazil, ACM**:** 669-670.

Zhang, Y. and P. Vines (2004). Detection and translation of OOV terms prior to query time. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Sheffield, United Kingdom, ACM**:** 524-525.

Zhang, Y. and P. Vines (2004). Using the web for automated translation extraction in cross-language information retrieval. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Sheffield, United Kingdom, ACM**:** 162-169.

Zhang, Y., P. Vines, et al. (2005). "Chinese OOV translation and post-translation query expansion in chinese-english cross-lingual information retrieval." ACM Transactions on Asian Language Information Processing (TALIP) **4**(2): 57-77.

Zhou, D., M. Truran, et al. (2008). "A Hybrid Technique for English-Chinese Cross Language Information Retrieval." **7**(2): 1-35.