OOV Term Translation, Context Information and Definition Extraction Based on OOV Term Type Prediction

Jian Qu, Akira Shimazu, and Le Minh Nguyen

School of Information Science, JAIST
Nomi, Ishikawa, Japan
{qujian,shimazu,nguyenml}@jaist.ac.jp

Abstract. Although there are many existing approaches for solving the OOV term translation problems, but existing approaches are not able to handle different types of OOV terms, especially hybrid translations, such as "Kenny-Caffey syndrome (Kenny-Caffey氏症候群)". We proposed a novel integrated ranking approach to consider the types of OOV terms before translating them. Thus, different types of OOV terms could be translated differently. Furthermore, the translations mined in other languages are also OOV terms, none of existing approaches offer the context information or definitions of the OOV terms. Users without special knowledge cannot easily understand meanings of the OOV terms. Our integrated ranking approach also extracts monolingual definitions and multilingual context information of OOV terms. Moreover, we propose a novel adaptive rules approach with Bayesian net and Adaboost for handling hybrid translations. Experiments show our approach performs better than existing approaches.

Keywords: Term translation, multilingual information retrieval.

1 Introduction

Out of vocabulary (OOV) terms are typically new terms that cannot be found in dictionaries. OOV terms can be classified into two groups; they are name type OOV terms such as personal names, place names, brands, etc. and technical type OOV terms such as new technical terms and new biomedical terms etc. Given an OOV term in source language, OOV term translation extraction aims to find the correct translation in target language.

Many existing approaches had explored various ways of finding translations for name type OOV terms in other languages. Biomedical type OOV terms, especially hybrid translations have received little attention in the past years. Hybrid translations use part target language and part source language. For example: a biomedical English OOV term "Kenny-Caffey syndrome" with its Chinese translation "Kenny-Caffey 症候群", "Kenny-Caffey" is source language and "症候群" is target language. Another fundamental problem for existing approaches is the translations obtained in other languages are also OOV terms. These translations provided little information to users

H. Isahara and K. Kanzaki (Eds.): JapTAL 2012, LNAI 7614, pp. 76–87, 2012.

without special knowledge. For example, "Mae West", its Chinese translation "梅蕙絲" does not offer any knowledge to users. Whether "梅蕙絲" is a company, a person or a Brand usually requires users to do further search.

In order to address the above problems, we propose to translate name type OOV terms and biomedical type OOV terms using different approaches. Thus, we will predict the type of OOV terms before translating them. We propose a novel integrated ranking approach to automatically predict the types of OOV terms. Then a novel adaptive rules approach together with supervised machine learning by Bayesian net with Adaboost is used for finding translations for biomedical type OOV terms, and we employ ranking list approach for finding translations for name type OOV terms. Furthermore our novel integrated ranking approach also extracts context information and definitions for name type and biomedical type OOV terms. For example, "Mae West", this approach would extracts its Chinese translation"梅蕙絲", multilingual context information in English (*American/ actress/ playwright/ screenwriter/ writer*) and Chinese (美国人/女演员/ 剧作家/ 编剧家/ 作家) and its monolingual definition (*Mae West (born Mary Jane West on August 17, 1893 – November 22, 1980) was an American actress, playwright, screenwriter and sex symbol whose entertainment career...*)[1].

The remaining of this paper is organized as follows: Section 2 introduces related works. Our approach is overviewed in Section 3. In Section 4, we discuss the experiments and the results. And finally in Section 5, we conclude this paper and discuss the future works.

2 Related Works

Many researches in the past have proved automatic web mining is the most efficient approach for translating OOV terms [1-3]. Most OOV terms have their correspondent human translations nearby on the Internet [2, 3]. The translations mined from the Internet are usually high quality and require low man power cost. Automatic mining include three major steps, they are: 1) web retrieval aims to collect the snippets containing the possible translations of the OOV term from the Internet; 2) translation extraction aims to find the boundary of the translations in the snippets; and 3) translation selection aims to choose the correct translation from the extracted translations.

Many researchers in the past have endeavored to solve the OOV term translation problems with a similar translation extraction approach from Zhang and Vines. [1, 2, 4-7]. Zhang and Vines extract up to 30 Chinese characters before and after the OOV term when English OOV term is found. Then they use brute force translation extraction to generate all possible substrings of the extracted Chinese characters. This approach has a very high recall but may generate many noises, and it is difficult to handle biomedical type OOV terms, especially hybrid translations, since only Chinese characters are extracted.

Translation selection can be generally categorized into statistical based approaches and machine learning approaches. Ranking list approach from Zhang and Vines is a typical statistical based approach [4]. The ranking list approach uses lengths and frequencies of translation candidates to select the correct translation. In addition to improve the ranking list approach, Cheng *et al.* suggested the Symmetrical Conditional

[1] Web retrieved definition.

Probability and context dependency (SCPCD) and Lu *et al*. used the Symmetrical Conditional Probability (SCP) for selecting the correct translation [2, 8]. Machine learning based approaches for translation selection was proposed by Tifin *et al*. [9]. Many machine learning approaches for translation selection utilize support vector machine (SVM) [7]. However, parameters are very important and expensive for SVM.

Existing approaches have a large drawback on hybrid translations, for example "Kenny-Caffey syndrome" (Human: Kenny-Caffey症候群) (existing approach: 症候群). If "症候群" is applied to CLIR, many disease documents unrelated to "Kenny-Caffey syndrome" will be retrieved, because many Chinese medical terms end with the term "症候群". Another fundamental problem of existing approach is the translations mined in other languages are also OOV terms, none of existing approaches offer the context information or definitions of OOV terms. Users without special knowledge cannot easily understand the detail meanings or the general ideas of OOV terms. It is very difficult for users without special knowledge to understand the meaning of a biomedical OOV term from only translations. Table 1 shows examples of translation results from our approach and existing approach. Seen from the table, our approach provides user the multilingual context information, which helps to explain the meaning of OOV terms.

Table 1. Examples of OOV term translations

OOV	Existing approach[4]	Our approach	
	Translation	Translation	Context information
Hereditary epidermolysis bullosa	先天性水皰症	先天性鬆懈水皰	skin disease/皮肤病
Leigh disease	萊氏症候群	萊氏症候群	rare neurometabolic disorder/罕见的神经代谢障碍

3 OOV Term Translation, Context Information and Definition Extraction Based on OOV Term Types

Existing approaches are very successful on name type OOV terms. However, they have many drawbacks on technical type OOV terms, especially hybrid translations from biomedical type OOV terms. In this paper, we developed a new adaptive rules approach for hybrid translations. However, this approach has some drawbacks on name type OOV terms.

In order to address the above problems, we propose to translate the OOV terms based on predication of types of OOV terms. Our approach takes into account a novel factor of different types of OOV terms. Thus different type OOV terms could be translated using different approaches. Existing ranking list approach is employed for translating name type OOV terms, and our novel adaptive rules approach is used for biomedical type OOV terms. A flow chart of our approach is shown in Fig. 1.

Seen from the figure, Our approach is developed into 5 steps, they are: 1) The snippet retrieval: documents (snippets) in both Chinese and English Languages containing OOV term and its possible translation, context information and definition are retrieved by

querying the English OOV terms over the Internet; 2) The snippet ranking and definition selection: English language snippets are ranked and OOV term definition are selected by our novel integrated ranking approach; 3) The multilingual context information extraction and OOV type predication: ontology tree is constructed from Word Net to extract the context information and predict the type of OOV terms; 4) The name type OOV term translation: name type OOV terms are translated by existing ranking list approach. 5) The biomedical type OOV term translation: biomedical type OOV terms are translated by our novel adaptive rules approach with Bayesian net and Adaboost.

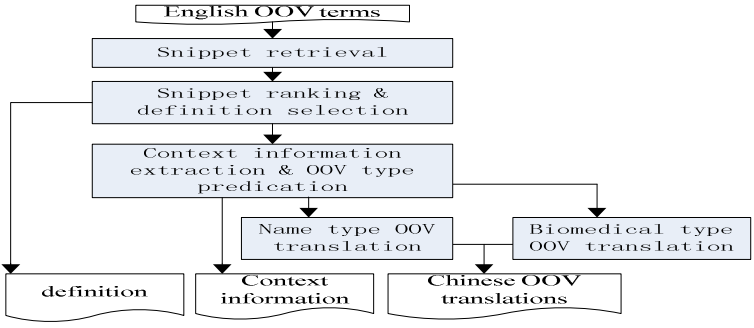

Fig. 1. Flow chart of our proposed approach

3.1 Snippet Retrieval

We feed English OOV terms to Bing Search API[12] and retrieve snippets from English language. We also retrieve snippets from Chinese language by limiting the retrieval language to Chinese. An example of English snippet containing the OOV term and its definition, and an example of Chinese snippet containing the OOV term and its translation are shown in Fig. 2.

English snippets:
Mae West - Wikipedia, the free encyclopedia (Title)
Mae West (born Mary Jane West on August 17, 1893 – November 22, 1980) was an American actress, playwright, screenwriter and sex symbol whose ... (Summary)
http://en.wikipedia.org/wiki/Mae_West (URL)

Chinese Snippets:
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)

Fig. 2. An example of web retrieved snippets

3.2 Snippet Ranking and Definition Selection

In order to predict the types of OOV terms, we need to firstly extract the definitions of OOV terms. According to our observation, many OOV terms have their definitions on the web retrieved snippets. There are three main factors which can help us to identify the snippet with definitions. They are: 1) The snippets with definition possess some patterns. For example, "Mae West ... was an American actress ..." Some verbs can be

[2] http://www.bing.com/toolbox/bingdeveloper

used to identify snippets with definitions. 2) Search Engine rank the webpage very carefully, high Search Engine rank can also led to snippets with definition. 3) Web-pages with definitions are usually well organized organization, government, or educational web-pages. Thus domain names of the web-pages can also led to snippets with definitions. Combine above three factors, we propose a novel integrated ranking approach to rank the snippets with definitions. This approach takes snippets retrieved from the Internet and a list of verbs that can identify the snippets with definitions. We consider the locations and co-occurrences between the verbs and the OOV terms. Then we combine the domain ranking and Search Engine ranking to select the snippets with definitions. Domain ranks are given as follows: gov/org/edu/int > com/pro/net/info/ > else.

The integrated ranking is developed as follows. Let Ss be the summaries of English snippets retrieved from the Internet, St be the titles of English snippets retrieved from the Internet, OOV be the source OOV terms, V be the verb list, SR be the rankings from Search Engine, and DR be the domain ranking. For each OOV term, if the OOV term and a verb in V are both found in the Ss, we give it a rank 1, if the OOV term is found in the St, the sub-word of the OOV term and a verb in V are found in the Ss, we give it a rank 2; if only sub-word of the OOV term is found in the Ss, we give it a rank 3; if no sub-word of the OOV term is found in the Ss, we give it a rank 4.

After the ranks are assigned to the snippets, for each OOV term, we select one snippet with the highest snippet rank, the highest Search Engine rank and the highest domain rank. For each selected snippet, we extract the summary of the snippet as the definition of the OOV term. A detailed algorithm of the integrated ranking is shown below.

Algorithm snippets ranking and selection

Input: Summaries of snippets retrieved from the Internet, Ss; Titles of snippets retrieved from the Internet, St; OOV terms, OOV; Verb list, V; Rankings from Search Engine, SR; Disambiguation noun list, N; Domain ranking list, DR.

Output: Snippets with ranking, SwR; Selected Snippets, SwD.

```
For each OOV do
  If (OOV and V found in Ss)then
SwR = 1,
  Else If (OOV found in St, sub-word OOV and V
  found in Ss)then
   SwR = 2,
   Else If (sub-word OOV found in Ss )then
     SwR = 3,
   Else
   SwR =4,
SwD = MaxRank(SwR) && MaxRank( SR) && MaxRank( DR)
End
```

An example of integrated ranking is shown in Table 2. As can be seen from this example, the snippet containing the correct definition of the OOV term "Mae West" gained a high rank.

Table 2. Ranking results of Snippets

URL	Title	Summary	Ranks(SWR, SR, DR)
http://en.wikipedia.org/wiki/Mae_West	Mae West - the free encyclopedia	Mae West (born Mary Jane West on August 17, 1893 – November 22, 1980) was an American actress, playwright, screenwriter and sex symbol whose entertainment career ...	1,1,1
http://www.imdb.com/name/nm0922213/	Mae West - IMDb	My Little Chickadee (1940) · Klondike Annie (1936). Mae West was born in Brooklyn, New York, to ...Soundtrack: I'm No Angel (1933) ...	1,2,2
http://www.youtube.com/watch?v=qVrfHXnUJFc	Mae West in I'm No Angel Trailer - YouTube	With Cary Grant in this 1933 comedy classic. Fortuneteller: I see a man in your life. Mae: What, only one?...	2,3,2

3.3 Multilingual Context Information Extraction and OOV Type Prediction

For each OOV term with its definition from the previous step, we extract the context information from the OOV term. We construct two ontology trees from Word Net to predict the type of OOV terms. Word net is a large English lexicon [10]. We use 7 Word Net super node noun terms (such as country, occupation, industry, etc) and extract their brief hyponyms to construct the ontology tree for name type OOV terms. Then we use 5 Word Net super node noun terms (such as illness,) and extract their brief hyponyms to construct the ontology tree for biomedical type OOV terms. We use the definition of each OOV term to search against the ontology trees. When any word in definition is found in the ontology trees, we extract such word as a context information of the OOV term. Then, we predict this OOV term to either name type or biomedical type according to the majority number of context information found in different ontology trees. Furthermore, these context information are then translated into Chinese by multilingual dictionary[3].

3.4 Translation Extraction and Selection for Name Type OOV Terms

After the types of OOV terms are predicted, the name type OOV terms are translated by existing ranking list approach [4]. For each OOV term, we select the top ranked candidates as the Chinese translation.

3.5 Translation Extraction and Selection for Biomedical Type OOV Terms

We propose a new adaptive rules approach for biomedical type OOV terms in order to handle hybrid translations. The translation extraction and selection for biomedical type OOV term is developed into 3 steps, they are: firstly, the translation extraction using novel adaptive rules approach; secondly, the feature extraction, and finally the translation selection using Bayesian net with Adaboost.

Translation Extraction for Biomedical Type OOV Terms

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,

[3] `http://www.oxfordlanguagedictionaries.com/Public/PublicHome.html`

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 approach.

The adaptive rules approach 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 extraction.

The adaptive rules approach 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 adapted 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 adapted 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 adapted 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 approach is explained below. An example of Re and Ar for OOV term "Kenny-Caffey syndrome" is shown in table 3.

Algorithm Translation 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: *Adapted 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
```

Table 3. An example of Re and Ar

#	Re	Ar
1	a-z/Chinese characters	Kenny-Caffey/Chinese characters
...	...	...

Feature Extraction

In this subsection, we extracted totally 19 different features from translation candidates. These features include: average distances, co-occurrence distance, term frequencies, symmetric conditional probability (SCP), modified association measures, lengths of OOV and translation, and length similarity. We describe the details of these features as follows.

Distances between OOV and translation

The closer a translation candidate to its source OOV term the more likely that translation is correct [2]. Some translations occur both 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 $Dist(e_i, c_i)$.

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 the above feature ($CwDist$) was proposed by Zhang *et al.* [7], they use the web retrieved page count instead of the $tf(c_i e_i)$.

Equation (1) shows the calculation of $CDist$, where $tf(c_i e_i)$ is the co-occur frequency between OOV and translation candidate.

Term frequencies

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.

Symmetrical Conditional Probability

Symmetrical Conditional Probability (SCP) [2, 8, 11] checks each alphabet, character and substring in the possible translation to determine whether this translation is a term or a sentence.

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 OOV terms $S(e_i)$, translations $S(c_i)$, the webpage count of OOV terms co-occur with translations $S(e_i \wedge c_i)$ and the webpage count of OOV terms 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 and sometimes hire human to eliminate the meaningless Chinese strings.

Support

$$Supp(e_i \rightarrow c_i) = S(e_i \wedge c_i) \tag{2}$$

Confidence

$$Conf(e_i \rightarrow c_i) = \frac{S(e_i \wedge c_i)}{S(e_i)} \tag{3}$$

Lift or Interestingness

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

Conviction

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

Equation (2) is the Support of association measure, equation (3) is the Confidence of the association measure, equation (4) is the Lift of the association measure, and equation (5) 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.

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 $D(|e_i|, |c_i|)$.

Length similarity
We also employ the length similarity ratio from Shi [12], it is a normalized length difference.

Translation Selection for Biomedical Type OOV Terms
In this section, we explain our candidate selection approach. It is developed into two parts, the statistical filter, and the Bayesian net translation selection with Adaboost.

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 Fig. 3.

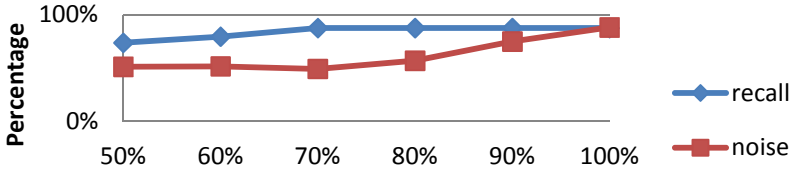

Fig. 3. Recalls of the filter

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 [13].

4 Experiments and Discussion

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

4.1 Data Collection

We collected English name type OOV terms from lists of top 500 companies[4], famous people[5], car brands[6], CPU code names[7]. Furthermore, we collected English biomedical terms from International Classification of Diseases, Functioning, and Disability version 9 (ICD9)[8]. We randomly selected 10% from each above list. Combining the above lists, we obtained a total of 203 OOV terms, 76 of them are name type OOV terms and rest are biomedical type OOV terms. The idea of this data collection is to test the flexibility and possibility of our approach on both name type and biomedical type OOV terms. In order to create a baseline, we hired two Chinese graduate students manually search the web/dictionary to find the correct OOV term definition, context information and Chinese translation.

We retrieved a total of 7,182 English snippets and 5,897 Chinese snippets after querying the above OOV terms to the web. Then we process these snippets with our integrated ranking approach.

4.2 Experimental Results for OOV Type Predication, Definition and Multilingual Context Information Extraction

Although the input were 203 OOV terms, only 201 were able to retrieve form the Internet. Our novel integrated ranking approach correctly predicted 75 name type OOV terms, and 118 biomedical type OOV terms. Furthermore, our integrated ranking approach also extracted the monolingual definition of the English OOV term; the monolingual context information of the English OOV term; and the Chinese translations of the context information. We compared each OOV term with the baseline to check for correctness. The detailed results are shown in Table 4.

Table 4. Experimental results for OOV type prediction, definition and context information extraction

Baselines	OOV type prediction			monolingual		multilingual
	precision	recall	accuracy	Definition(accuracy)	context information(accuracy)	Chinese context information(accuracy)
All OOV(201)	95.31%	96.54%	96.02%	170(84.58%)	181(90.05%)	179(89.05%)
Name type OOV(76)	91.46%	98.68%	96.02%	74(97.37%)	75(98.68%)	74(97.37%)
Biomedical type OOV(125)	99.16%	94.40%	96.02%	96(78.80%)	106(84.80%)	105(84.00%)

4.3 Experimental Setup for OOV Term Translation

After we predicted the types of OOV terms, the name type OOV terms are translated using ranking list approach and biomedical type OOV terms are translated using our novel adaptive rules and Bayesian net with Ada boost approach. RapidMiner [13] is used for

[4] http://money.cnn.com/magazines/fortune/fortune500/2011/index.html
[5] http://www.selfcreation.com/creation/famous_people.htm
[6] http://wiki.answers.com/Q/List_of_car_brands
[7] http://www.cpubenchmark.net/cpu_list.php
[8] http://www.cdc.gov/nchs/icd/icd9.html

machine learning since it details out each results on the learning process. We used 10-fold cross validation to experiment on the annotated data collection. Although some OOV terms have few correct translations, we only select one for the final evaluation.

To compare our approach with existing approaches, we tested the same data set with the Pat-tree and SVM approach from Zhang *et al*. [7]. We also tested the ranking list approach from Zhang & Vines [4].

4.4 Experimental Results for OOV Term Translation

Table 5 shows the experimental results for OOV term translation. We can see that ranking list approach from Zhang & Vines achieved accuracies of 77.61% and 67.66% in translation extraction and translation selection respectively. While Pat-tree and SVM approach from Zhang *et al*. gained accuracies of 78.61% and 71.14% in translation extraction and translation selection respectively. Our proposed approach is significantly better than existing approaches. We gained accuracies of 93.04% and 87.56% in translation extraction and translation selection respectively.

Table 5. Comparison of our proposed approach with existing approaches

Approaches	Transliterat-able OOVs	Correct translation mined (Name type)(biomedical type)	Correct translation mined in OOV terms(Name type)(biomedical type){accuracy}	Correct translation selected in OOV terms(Name type)(biomedical type){accuracy}
Our approach	201	212(79)(125)	187(74)(113) {93.04%}	176(74)(102) {87.56%}
Pat-tree SVM ap-proach		178(79)(99)	158(73)(85) {78.61%}	143(71)(72) {71.14%}
Ranking list approach		179(81)(98)	156(74)(82) {77.61%}	136(74)(62) {67.66%}

4.5 Discussions

Our approach performed better because we consider the types of OOV terms. Different approaches were used for different types of OOV terms. Furthermore, our adaptive rules approach can extract many translations where existing approaches failed to extract, mostly because we considered the existence of the hybrid translations. Moreover, our approach extracts the monolingual information and the multilingual context information for the OOV term. However, some drawbacks are within our integrated ranking approach. If Search Engine, our algorithm and domain ranking all ranked a wrong snippet highly relevant to an OOV term, we may get an error. For example, query term #144 "Fucosidosis"(a biomedical OOV term), however the snippet with high rank for "Fucosisdosis" contains definition for a person named "Fucosisdosis".

5 Conclusion

OOV term has always been a problem for natural language processing, especially for information retrieval. Although there are many existing approaches for solving the OOV term translation problems, but existing approaches are not able to handle different types of OOV terms. Furthermore, the translations mined in other languages are also OOV terms, none of existing approaches offer the context information or

definition of the OOV terms. Users without special knowledge cannot easily understand the detail meanings or the general ideas of the OOV terms. We proposed an integrated ranking approach for predicting the types of OOV terms and extracting the monolingual definitions and the multilingual context information. Moreover, we propose a novel adaptive rules approach with Bayesian net and Adaboost for handling hybrid translations. We evaluate our approach with both name type and biomedical type OOV terms. Our approach achieved high accuracies of 96.02% for OOV type prediction, 84.58% for monolingual definition extraction, and 89.06% for multilingual context information extraction. Our approach also achieved high accuracies of 93.04% in translation extraction and 87.56% in translation selection for OOV term translation. In future, we will develop better translation extraction approach, and improve our OOV context information extraction approach.

References

1. Lu, W.-H., Chien, L.-F., Lee, H.-J.: Anchor text mining for translation of Web queries: A transitive translation approach. ACM Trans. Inf. Syst. 22(2), 242–269 (2004)
2. Cheng, P.-J., et al.: Translating unknown queries with web corpora for cross-language information retrieval. In: ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 146–153. ACM, Sheffield (2004)
3. Zhang, Y., Huang, F., Vogel, S.: Mining translations of OOV terms from the web through cross-lingual query expansion. In: ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 669–670. ACM, Salvador (2005)
4. Zhang, Y., Vines, P.: Using the web for automated translation extraction in cross-language information retrieval. In: ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 162–169. ACM, Sheffield (2004)
5. Zhang, Y., Vines, P.: Detection and translation of OOV terms prior to query time. In: ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 524–525. ACM, Sheffield (2004)
6. Zhang, Y., Vines, P., Zobel, J.: 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 (2005)
7. Zhang, Y., Wang, Y., Xue, X.: English-Chinese bi-directional OOV translation based on web mining and supervised learning. In: ACL-IJCNLP 2009 Conference Short Papers, pp. 129–132. Association for Computational Linguistics, Suntec (2009)
8. Lu, C., Xu, Y., Geva, S.: Translation disambiguation in web-based translation extraction for English-Chinese CLIR. In: ACM Symposium on Applied Computing, pp. 819–823. ACM, Seoul (2007)
9. Tiffin, N., et al.: Integration of text and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res. 33, 1544–1552 (2005)
10. Fellbaum, C.: WordNet An Electronic Lexical Database (1998)
11. Ferreira da Silva, J., Dias, G., Guilloré, S., Pereira Lopes, J.G.: Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In: Barahona, P., Alferes, J.J. (eds.) EPIA 1999. LNCS (LNAI), vol. 1695, pp. 113–132. Springer, Heidelberg (1999)
12. Shi, L.: Mining OOV Translations from Mixed-Language Web Pages for Cross Language Information Retrieval. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 471–482. Springer, Heidelberg (2010)
13. Rapidminer, Rapidminer data mining tool (2009)