Web Translation of English Medical OOV Terms to Chinese with Data Mining Approach

Jian Qu, Thanaruk Theeramunkong, Cholwich Nattee, and Pakinee Aimmanee
School of Information and Computer Technology
Sirindhorn International Institute of Technology, Thammasat University
131 Moo5, Tiwanont Rd., Bangkadi, Muang, Pathum Thani, Thailand 12000
Email: {jian.qu}@studentmail.siit.tu.ac.th, {thanaruk, cholwich, pakinee}@siit.tu.ac.th

Abstract

It has always been so difficult to successfully translate OOV terms form one language to others, especially for those languages without clear word boundaries, such like Chinese. Some Chinese translations of English medical OOV terms are not perfectly translated, most of which involved non-Chinese character in the translation. Since all existing approach for translating English OOV term into Chinese handles the retrieved snippets from the Internet by extracting only Chinese characters, as a result the non-Chinese characters part of a Chinese translation would be lost. We propose an rule based method by considering each input English OOV term and automatically adjust the candidate generating system accordingly, unlike most existing OOV term translation approaches, we use a data mining approach rather than a frequently used statistic based approach to select the final translation candidate. By testing our approach with two of the most difficult English medial OOV term sets ICD9-CM and ICD9, it outperforms the existing approach with a precision of 89.98% and is able to handle the Chinese translation that includes non-Chinese characters.

Keywords: out of vocabulary (OOV), association measures, sematic conditional probability(SCP), decision tree.

1. Introduction

Today information spreads fast through our world mainly in several different ways, the newspaper, television and the Internet. Neither newspaper nor television has the speed and availability of the Internet, sometimes called our globally interconnected information infrastructure. It offers an easy and fast access to information worldwide with no boundaries. However, it limits users to their own abilities, for persons who read only English see only the English web; a reader who understands only Chinese sees only the Chinese web; and someone who reads only Thai, sees only the Thai web.

Cross-language information retrieval (CLIR) assists a user to issue queries in a source language to find information written in target language(s). CLIR has many useful applications. For example, a user on the Internet might want to issue a query in one language to find documents in many other languages. Lots of research has been done on CLIR, but still today, one of the best bi-lingual translation software for English to Chinese/Chinese to English -"King Soft" is only able to reach 40% to 50% accuracy when doing a newspaper article translation.

A major problem of the CLIR is the Out of Vocabulary (OOV) terms, which are

typically new terms from current affairs, such as person names, location names, new technical terms and translated words. Recent research has approached English to Chinese OOV term translation, most of which focuses on name entity OOV terms [1, 2, 3]. Name entity OOV terms are typically famous person names, location names, and brand names, which occur frequently in newspaper articles. However, translation of some special medical OOV terms are also needed because of the recent growing of cross country patients. Researchers incorrectly translated globally spread dangerous diseases at growing multinational hospitals. Whenever a new sickness is discovered, usually it is initially given a name in English. In most non-English speaking countries, the special medical terms are too difficult to be translated, so the original English terms are used. But the original English terms does not give any meaning for foreigners if it cannot be found in a bilingual dictionary. Therefore, those medical OOV terms need to be translated either by manually adding translations to the bilingual dictionary or by web retrieval of translations. While web retrieve takes advantage of the fast updating Internet.

It has always been difficult to successfully translate OOV terms from one language to other, especially for those languages without clear word boundaries, such as Chinese. Most existing system of translating English OOV terms to Chinese are focused on name entity OOV terms, so a standard process is to do the translation without segmentation, because the Chinese translation are OOV terms which do not exist in a monolingual dictionary. Segmentation can usually segment into smaller sequences of characters or individual characters [1, 2, 3]. For those person names or location names, the translation in Chinese tends to use rarely occurring characters. This makes segmentation useless. For most medical OOV terms, by observation we can suggest certain parts in the Chinese translation that can be segmented by a monolingual dictionary [4], which offer us a better window size for how many Chinese characters to select.

According to our observations, some English OOV terms are not perfectly translated into Chinese. Those Chinese translations tend to include non-Chinese characters. At the present, to our knowledge, the existing English to Chinese OOV term translation system cannot handle the translation if it has non-Chinese characters as a part of the translation. Some examples of Chinese translations of English medical OOV term are shown in Table 1.

Table 1 Few examples of Chinese translation of English medical OOV term

English OOV term	Correct Chinese translation	Chinese translation by existing system[1]
Kenny-Caffey syndrome	Kenny-Caffey 氏症候群	氏症候群
DiGeorge's syndrome	DiGeorge's症候群	症候群
α1- Antitrypsin deficiency	α1-抗胰蛋白□缺乏症	抗胰蛋白□缺乏症
3-Hydroxy-3-methyl-glutaric acidemia	3-□基-3-甲基戊二酸血症	甲基戊二酸血症
GM1/GM2 gangliosidosis	GM1/GM2神經節甘脂儲積症	神經節甘脂儲積症
Huntington's chorea	亨汀頓氏舞蹈症	亨汀頓氏舞蹈症

In this approach, we propose to use a rule based candidate generation system to handle the non-Chinese character problem of the Chinese translation. We select the Chinese translations using a machine learning system by considering the modified association measures between the OOV term and its translation candidates, exact distance of each Chinese translation candidates, candidates co-occurrence frequencies, chi-square, and semantic conditional probability as features.

2. Related Work

Large numbers of paired text in different languages are available on the Internet, most of which are parts from technical research papers, organizational or government web sites. This resource is perfect for OOV terms translation, simply because the authors of these pages tend to put in translations for the OOV terms. For example, when a new medical term in English is written in a research paper online in a Chinese language web site, the Chinese translation for this English term usually exists in front or behind the original English term. Zhang and Vines [5] state that if English terms occur on Chinese web pages, and if they also exists within brackets, they have a very high probability to be the translation of the nearby Chinese term. Existing research has proposed methods to retrieve those web snippets via Yahoo or Google.

Existing OOV term translation systems usually consider the co-occurrence frequencies and the length of the extracted terms for the OOV translation [1], which suffers from partial answer of a translation, multiple translated candidates, and require humans to select the correct translation [1, 3]. Many different medical terms occur together in one sentence in web text, and the Chinese translations for different medical terms are not too different from each other. If we simply apply the most referred English OOV term to Chinese translation method from Zhang and Vines [1], it would pick many wrong translations. Besides that, their method picks three to four possible Chinese translations and requires humans to select which one is correct. (query 9: Carlsberg 嘉士伯/节点类型节点类型列子/的节点类型节点类型列) [1] This make sense when we are looking for Name entity OOV terms, but for Medical OOV terms it picks up several different diseases, human without medical background cannot tell which one is correct.

Some Chinese translations of English medical OOV terms are not perfectly translated, most of which include non-Chinese characters. Since all existing systems for translating English OOV terms into Chinese handle the retrieved snippets from the Internet by extracting only Chinese characters [1, 3, 5, 6], Chinese translations with non-Chinese characters are lost. This problem usually does not occur in name entity OOV terms but it is a major issue for technical and medical type OOV terms. In this paper, we approach the Chinese translations of English medical OOV terms with a novel rule based candidate generation system. The extracting strings from the retrieved snippets can not only include Chinese but also English, number and symbols. Word segmentation system together with our candidate generation system is used to detect the boundary of how many words or characters are to be included. The distance between an English OOV term and its Chinese translation candidates is necessary, because the closer the distance, the higher the possibility to be the correct translation. Association measures proved to be useful in our previous work. [7]. Semantic conditional probability (SCP) [3, 6, 8] can help us to clarify which extracted candidates are more likely to be a word or a sentence. Existing CLIR research suggests that machine learning approaches outperform rule based approaches [9]. In order to consider the distance, frequency, SCP and association measures all together as features, we apply a decision tree to select the correct translation candidates.

3. Our Approach

In this section we discuss the automatic extraction of Chinese translations for English medical OOV terms. We apply eight steps to extract the correct Chinese

translation of the English medical OOV terms, The steps are as follows:

1) Extract English medical OOV terms from the web;
2) Segment the web snippets;
3) Determine the boundaries of the web retrieved snippets;
4) Extract Chinese translation candidate for the English OOV terms;
5) Extract features of each Chinese translation candidates;
6) Filter the Chinese translation candidate;
7) Generate candidate selection model by a decision tree;
8) Select the Chinese translation candidates.

A flow chart of this work is shown in Fig. 1.

3.1 Extract English medical OOV terms from the web

We feed the English medical OOV terms to Yahoo API[1] with two limits. The first limit is, we focus on traditional Chinese web pages since results in our previous experiments shows that traditional Chinese web site provide more translation pairs for medical terms. The second limit is, we only retrieve the complete English medical OOV terms, unlike name entity OOV terms, some medical OOV term contains partially dictionary translatable words, such as " α 1-antitrypsin deficiency", if we simply query it through the web, we can get the translation for "antitrypsin" and "deficiency" which would add noise to our candidates. An example of snippet containing English OOV term and its translation is shown in Fig. 2.

HTML tags are removed and the obtained web texts are separated into three different fields in the database, the URL, the Title and the Summary.

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 ofα1-antitrypsin deficiency

3.2 Segmentation the web snippets

Chinese word segmentation groups a few Chinese characters together if that pattern is found in a segmentation dictionary. We ran the data through

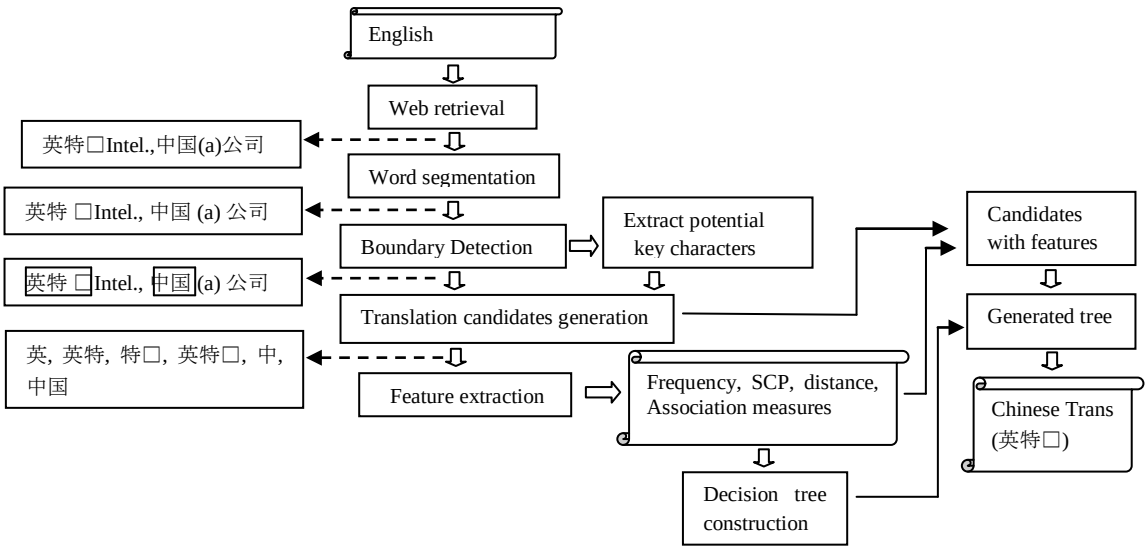

Figure 1 Flow chart of English OOV term translating into Chinese

a dictionary based Chinese segmentation system. Dictionaries have been proved very useful during CLIR [6, 10, 11]. The monolingual dictionary contains 112,492 traditional Chinese word phrases. Since partial English medical OOV terms are bilingually translatable, translation would result in some Chinese word phrases and others into single characters. The segmentation system can separate the punctuation away from Chinese characters, which is fundamental for our rule based candidate generation system, shown in Fig. 3 and 4. We do not consider any retrieved snippets that do not contain any Chinese characters.

2009年1月10日 ... α1-抗胰蛋白□缺乏症(α1-antitrypsin deficiency)是以□儿期出□胆汁 ... 的糖蛋白，在化学□成上与正常α1-AT 的区□是缺乏唾液酸基和糖基。

Figure 3 An example of the original data

2009 年 1 月 10 日 ... α1- 抗 胰 蛋白 □ 缺乏症 (α1-antitrypsin deficiency) 是 以 □儿 期 出□ 胆汁 ... 的 糖 蛋白，在化学 □成 上 与 正常 α1-AT 的 区□ 是 缺乏 唾 液 酸 基 和 糖基。

Figure 4 The data after segmentation

3.3 Determine the boundaries of the web retrieved snippets

The distance between the Chinese translation candidates and the English medical OOV term is important in this work, because the closer the Chinese word phrase the higher possibility of the correct translation [3, 12].

Since we would like to handle Chinese translation that includes non-Chinese characters, we simply include all English, numbers and symbols for each translation candidate. This adds lots of noise and the final translation can be even worse than just extracted Chinese characters. For each English OOV term, we check this English OOV term for its potential key characters first, and we classify them into four different rules.

1. If the English OOV term itself contains just English, then the potential key characters would be sub English words and any Chinese characters;

2. If the English OOV term itself contains symbols, then the potential key characters would be sub English words, any Chinese characters, and the same symbols;

3. If the English OOV term itself contains numbers, then the potential key characters would be sub English words, any Chinese characters and the same numbers;

4. If the English OOV term itself contains numbers and symbols, then the potential key characters would be sub English words, any Chinese characters, and the same numbers and symbols.

For each snippet containing English OOV term, we detect the boundaries of the string in front and behind the English OOV term according to the following steps.

1. We scan for each character and word to check for the first Chinese characters within a searching distance of 8 characters or words because we discovered that a searching distance equal to or more than 8 provides the best recall. Recalls of different searching distances are shown in Fig 5.

2. If any Chinese character is found within the searching distance, we call that Chinese character the initial Chinese character. Otherwise the search stops.

3. We extract all continuous potential key characters from both sides of the initial Chinese character.

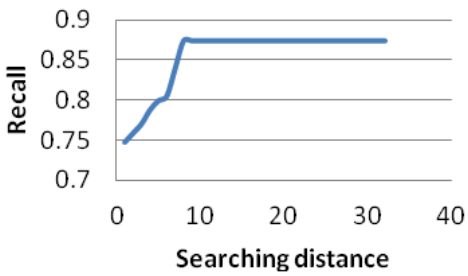

Figure 5 Recalls for different searching distances

An example of boundary detection is shown in Fig. 6. In this example, the English OOV term is “α 1- Antitrypsin deficiency”, we found the initial Chinese character in front of the English OOV term is “病”, and the initial Chinese character behind the English OOV term is “抗”. Then, we extract for any continuous potential key characters from both sides of these initial Chinese characters. The extractions stopped at the “.” because “.” is not a potential key character.

3.4 Extract Chinese translation candidates

Since the previous step only identifies the boundary of the translation candidates, all possible patterns in both the front string and back string are considered in this step, for example, a string with five characters/words “Robinow 氏症候群”, generates substrings as Robinow, 氏, 症, 候, 群, Robinow 氏, Robinow 氏症, Robinow 氏 症候, 氏症, 氏症候, 氏症候群, 症候, 症候群, 候群and Robinow 氏症候群.

We focus on medical OOV terms, for the translation in Chinese, they usually end with certain keywords [13], such as sickness " 症" or "病", syndrome "候群", disorder "異常" or "缺陷". We generated a group of keywords for Chinese medical terms from a publicly available list of diseases.

By using all possible patterns and the keywords we generate up the possible Chinese candidates.

3.5 Extract features of each Chinese translation candidates

In this section, we describe the details of all features we extracted from each Chinese translation candidate. These features include: frequency, average distance, SCP, modified association measures, and chi-square.

1) Frequency

Frequency represents a very important statistical feature for the retrieved Chinese translation candidates. The more a Chinese translation co-occur with an English OOV term, the more likely it is a correct translation. Collecting the frequency of the Chinese translation candidates is just counting how many times those translation candidates occur in our web retrieved snippets. There are however three different frequencies of each Chinese translation candidates.

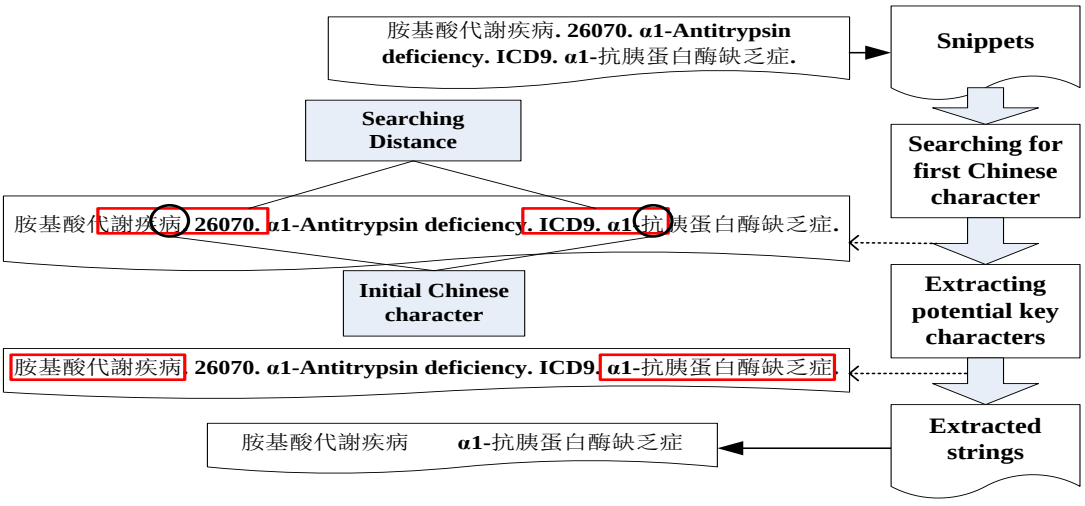

Figure 6 Boundaries detection

The web retrieved snippets are separated into two different parts, the part before the English OOV term, and the part after it. So we extract the front frequency, back frequency, and the total frequency for each Chinese translation candidates.

2) Front and back average distance of Chinese translation candidates

The closer a Chinese translation candidate to its English OOV term, the more likely that a Chinese translation is correct [3, 12]. Calculating the distance for each Chinese translation candidates is just counting how many words or characters that Chinese translation candidate is away from the English OOV term in the web retrieved snippets. Note that there are strings in front of the English OOV term and after the English OOV term, and we take into consideration the distance for each candidate. This candidate can be in the front part or the back part, and we cannot just put their distance together, because the front distance counts from the last character/word of the candidate, but the back distance counts from the first character/word of the candidates. So the distances are separated for both front and back. Distance is an average on each candidate, the back distance only contains the average for the back part, while the front frequency only contains the average for the front part, shown in Fig. 7.

Citrullinemia. 中文是瓜胺酸血症 .

Citrullinemia. 瓜胺酸血症 .

Candidates	Average distances
瓜胺酸血症	4
胺酸血症	5
酸血症	6

Figure 7 Example of distance

3) Symmetrical Conditional Probability

Symmetrical Conditional Probability (SCP) [3, 6, 8] checks each character and substring in the possible Chinese translation. Calculating the frequency of each substring in the corpus and comparing them to the frequency of the Chinese translation, the substring of the Chinese translation occur less often in the corpus than it occurs 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.

We consider each Chinese substring in a Chinese translation, for example a Chinese translation "Kearns Sayre 氏症候群 ", so Kearns, Sayre, 氏, 症, 候, 群, Kearns Sayre, Kearns Sayre 氏, Kearns Sayre 氏症, Kearns Sayre 氏症 候, Kearns Sayre 氏症候群, Sayre 氏, Sayre 氏症, Sayre 氏症候, Sayre 氏症候群, 氏症, 氏症候, 氏症候群, 症候, 症候群 and 候群 are all substrings for the translation. If all substrings occur in the corpus at the same time as the frequency of the word, then the SCP for this word is maximum, which equals one. The more times the substring occurs, the lower the SCP score.

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

Equation (1) shows the calculation of SCP,where $(c_1...c_n)$ is any possible Chinese translation candidates generated, n is the number of how many characters this Chinese translation candidate has, $f(c_1...c_n)$ is the frequency of that candidate, and $f(c_1...c_i)$ or $f(c_{i+1}...c_n)$ is the frequency of any substring of the Chinese translation candidate.

4) Modified Association Measures

Association measures were proved to be useful in our previous work and also other text mining research [7, 14]. We

applied them to find the relationship of each Chinese translation candidates to its associated English OOV term. The raw data to compute the association measures can be found by counting the page number returned by querying English OOV terms or Chinese translation candidates or English OOV terms together with Chinese translation candidates on the web. However, traditional association measures require the total amount of data, which is the total number of pages on the Internet, if the data are number of pages. We modified the association measure, so they do not require the total number of pages on the Internet.

Support

Support is an undirected measure that finds the ratio when e_i and c_i occur together in the same data set, and it is computed as follows.

$$Supp(e_i \rightarrow c_i) = S(e_i \wedge c_i) \qquad (2)$$

Confidence

Confidence is a directed measure for the ratio that e_i occurs when c_i occurs:

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

Lift or Interestingness

Lift or interestingness is the correlation between e_i and c_i. It tests on two hypnosis. First the occurrence of e_i is independent of the occurrence of c_i or second e_i and c_i are dependant and correlated in the same data set. It is computed as follow.

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

Conviction

Conviction represents the ratio of the expected frequency that e_i occurs without c_i, and it is computed as follows:

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

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 association measure. e_i is the English OOV term, c_i is the generated Chinese translations and S(e_i) is the number of pages returned by the search engine when e_i is submitted as a query.

5) Chi-square

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

$$\chi^2(e_i, c_i) = \frac{N \times (a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)} \qquad (6)$$

where the meaning of each variable is explained as follows:

e_i	=	source OOV term
c_i	=	translation candidates
a	=	$S(e_i \wedge c_i)$
b	=	$S(e_i \wedge \neg c_i)$
c	=	$S(c_i \wedge \neg e_i)$
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

3.6 Filter the Chinese translation candidate

Since there are too many Chinese translation candidates to process forward, a statistical ways to filter the Chinese translation candidates is performed. In each set of Chinese translation candidates from an English OOV term we take the candidates with top 70% from the frequency rank, since it provide the best recall and the lowest noise, where noise represents the number of wrong translations. Recalls using the top 70% frequency rank are shown in Fig. 8.

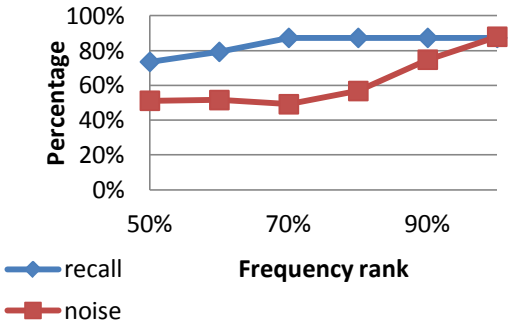

Figure 8 Recalls of frequency rank

However, the frequency filter still offers us more than 100 Chinese translation candidates for each English OOV term. We can reduce this number by simply taking the lowest distance in each set of translation candidates, because the closer the translation candidate is to the English OOV term, the higher the possibility that the translation is correct.

3.7 Generate candidate selection model by decision tree

Data mining tools have been proven to be very useful in CLIR [8]. Research suggests that data mining outperforms rule based candidate selection approach in multiple ways. Our features include modified association measures, front and back average distances, frequency, and SCP. Applying them together mostly results in a rule set. A decision tree [15, 16] is reasonable to apply to those features. A decision tree requires at least more than one class to run the 10-fold cross validation. We can manually check the results from Chinese translation candidates with the list of translation provided by the Taiwanese government. The candidates that are correct can be classed in a correct class. The candidates that are incorrect can be classed in a wrong class. Then we use the C4.5 algorithm to run the translation candidates to generate the model tree.

3.8 Select the Chinese translation candidates

The selection process is straightforward, based on the model tree. We run the translation candidates on the tree and pick up the ones that are correct, according to the tree.

4. Experiments and Results

In this section, we describe the experiments and results for extracting Chinese translations from English medical OOV term.

4.1 Experiment setup

1) Documents collection

We collected two data sets of English OOV terms from different sources. The first data set is the list of English special medical OOV term from Centers for disease control, R.O.C. (Taiwan)[1]. The list is called: "公告罕见疾病名單暨 ICD-9-CM 編碼一覽表"[2]. It contains totally 184 ICD9 CM special medical terms. They are presented in both English and Chinese. The English list will be used as our input, and we compare our experimental results with the Chinese list. The second data set is the English medical terms from Classification of Diseases, Functioning, and Disability[3]. The list is called "Inter- national Classification of Diseases, Ninth Revision (ICD-9)". We extract any English medical terms that can be found on the Chinese website according to Yahoo API. A total of 240 English special medical terms were extracted, which is different from the first set. We manually query each English medical term on the web and com- pare it to the published resource of Taiwan CDC[4], to find the correct Chinese translations for comparison with our experimental results later.

2) Segmentation tool

We used DEDE Chinese segmentation system[5] due to its ability to handle traditional Chinese characters.

3) Keywords list extraction

A list of possible keywords for Chinese special medical term was extracted from the Taiwan Government web services for genetic disease[6].

4) Data mining & error analysis tool

Rapid Miner[7] is used for doing the data mining and error analysis since it details each result in the learning process.

4.2 Experiments

We first ran data set 1, data set 2 and a combination of data set 1 and data set 2 of English special OOV terms with the length and co- occurrence frequency rule, and selected the Chinese translation based on the method purposed by Zhang [1] in his paper. Only one unique answer for each OOV term was selected. The result is for the final comparison with our experiment, to determine whether our approach is better than the existing ones or not.

We ran data set 1, data set 2 and a combination of data set 1 and data set 2 of English special OOV terms with our rule based candidate generation and decision tree approach. In order to make the decision tree, we manually classified the output data into two classes. The correct class and the wrong class. Then, the decision tree is applied with 10-fold cross validation.

4.3 Experiment results and Discussion

We have to separate the results in this part, since there are results from the candidate retrieval, results from candidate generation, and results from decision tree selection process. Some translations can be lost from the retrieved snippets, since there is no Chinese translation for that specific English OOV term on the Internet. This problem is caused by the translations of those English OOV terms in Chinese that exist in a picture format and the search engine cannot find them easily. We have to take only the web retrievable English OOV terms to consider as the base line. For the candidate generation part, some English OOV terms may have the correct Chinese translation, but may be lost due to too many incorrect translations in the snippets. Here we study how many correct ones are manually retrievable and how many are retrieved by our system. In the decision tree, we take the data classified into correct class and wrong class to run the 10-fold cross validation. Since there are many features used in the decision tree, we check the performance of all combinations of features. Finally, we choose the best combination to generate the model tree and select the Chinese translations.

We have the correct ICD9-CM Chinese translations provided by the Taiwanese government and correct ICD9 Chinese translations retrieved from the Internet. We evaluate the correctness, by comparing the results we have with the correct Chinese translations. If the result is semantically the same as the Chinese list, we say it is a correct match. If a result is semantically different from the given list, we call it a mismatch.

For our experiment, although the input is 424 English OOV terms, the web retrievable OOV terms are only 419 terms. Among that, 409 have the correct translation in the retrieved snippets if checked manually. Details are shown in Fig. 9.

Because some English OOV terms may have more than one correct or wrong translation, there are 742 pairs of Chinese translation candidates with their English OOV terms to put into a decision tree. The decision tree on all features and the overall English OOV term translations are shown in Fig. 10. Since there are different features in

the decision tree, some of them may conflict with each other, so we evaluate each feature by running the decision tree for all possible combinations of the features, except frequency and distance. The frequency and distance were used in the candidate filtering process and the filtered candidates usually have high frequency and low distance. A list of performances for different combinations is shown in Table 2 .

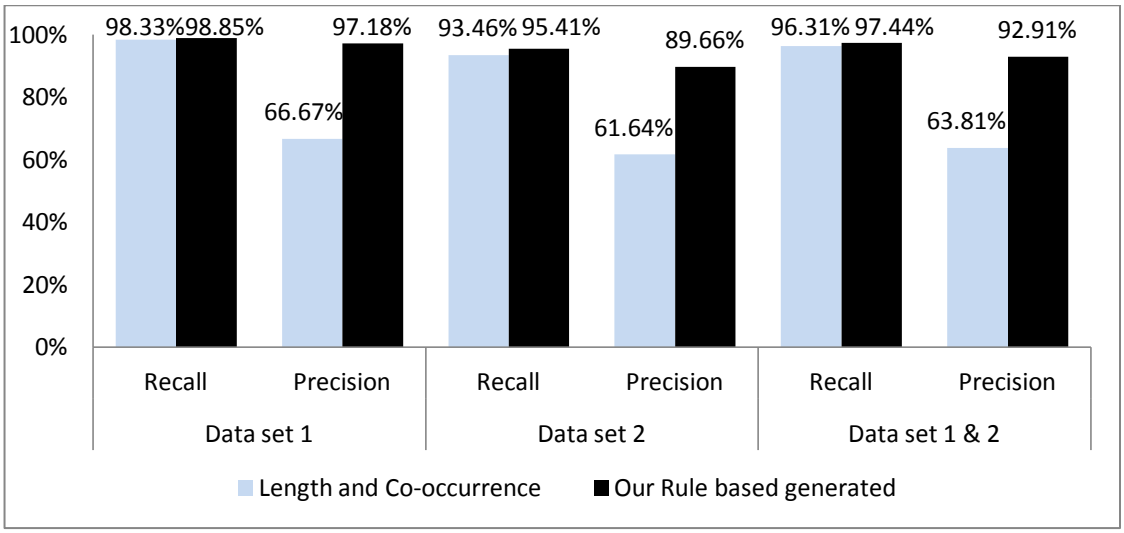

Figure 9 Comparison of candidate generation for Length Co-occurrence from Y. Zhang and our rule based approach

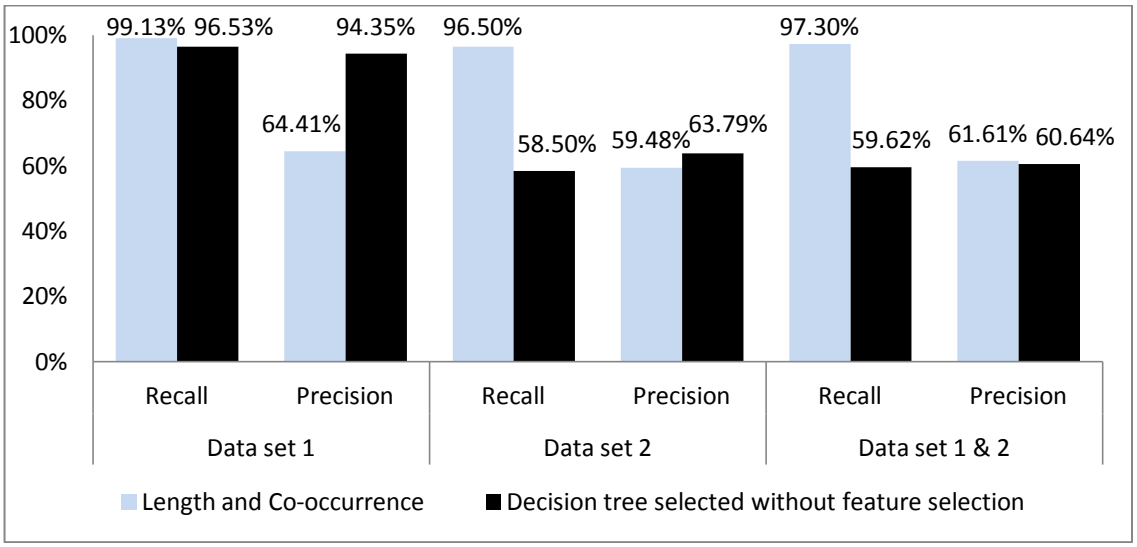

Figure 10 Comparison of overall English translating by length and co-occurrence from Y. Zhang with our decision tree approach without feature selection

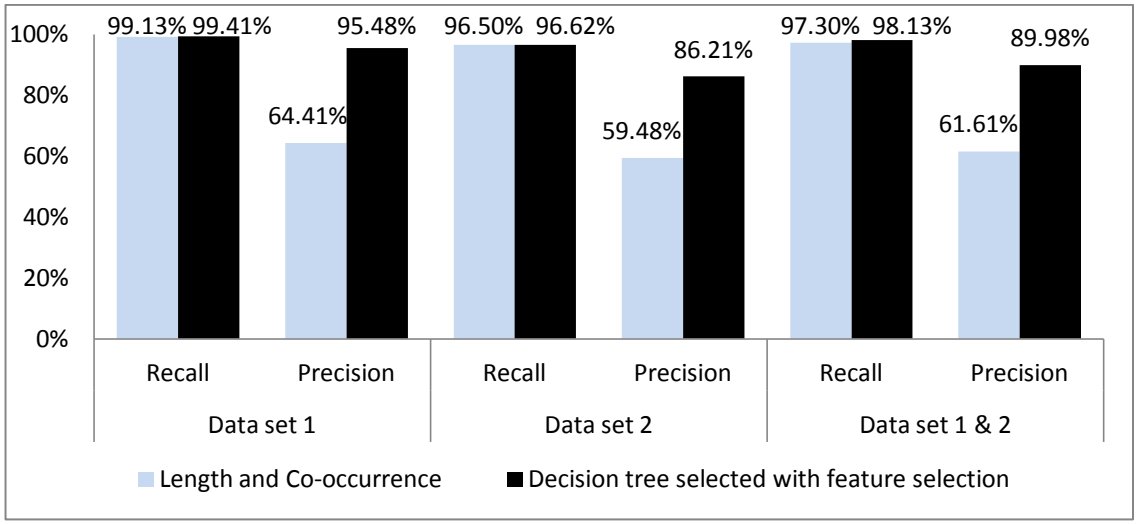

Figure 11 Comparison of overall English translating by length and co-occurrence from Y. Zhang with our decision tree approach with the best features from feature selection

According to our experiments, the best candidate selection result is by using the features such as lift together with frequency, and distance. The accuracy is up to 86.7%. In the top range of Table 2, where the high performances are shown, all include our modified association measures as features. When Chi-square is included, the accuracy drops to 59.43%. This is because the Chi square is calculated from retrieved snippets, not the search engine returned number of pages. Details are shown in Table. 2.

Finally we compute our candidate selection result in our candidates generation system and check how many English OOV term are correctly translated. We compare it with Zhang's length Co-occurrence result. We have a recall of 98.13% and a much higher precision of 89.98%, shown in Fig.11.

4.4 Error analysis

Those wrongly classified instances in the decision tree have two conditions. First, a correct instance can be classified into a wrong class; second, a wrong instance can be classified into correct class. The decision tree tried to find a general rule for all translation pairs of English medical OOV term, regardless of some translation pairs occurring more often and some occurring less often. Some translation pairs with variable bias results will be classified wrongly.

When using different attributes association measures always play a very important role. When we calculate the association measures, we get the number of returned pages from the Inter- net as raw data. This number only checks the co-occurrence of the input English OOV term and the Chinese translation candidates. It did not check whether these Chinese candidates may be very far away from the input English OOV term, because they all result in the same returned page number. Even though those precautions were taken during the candidates filtering process by selecting large frequency and low distance, few of them get past the filter and cause this trouble. When a wrong instance is classified into a correct class, another disease translation may co-occur much more often with the English OOV term than the correct ones, such as queried term 160 (Tyrosinemial 酪胺酸血症), but the obtained Chinese translation is "胺基酸代謝疾病", which is the translation

Table 2 Experiments for all combinations of features for candidate selection based on Chinese translation candidates

Combinations	TP	FP	FN	TN	Precision	Recall	Accuracy	F-measure
B	400	41	57	244	90.70%	87.53%	86.79%	89.09%
BE	399	42	58	243	90.48%	87.31%	86.52%	88.86%
ABD	398	43	57	244	90.25%	87.47%	86.52%	88.84%
ABCD,ABCDE	397	44	56	245	90.02%	87.64%	86.52%	88.81%
ACD,ACDE	393	48	51	250	89.12%	88.51%	86.66%	88.81%
AB	398	43	58	243	90.25%	87.28%	86.39%	88.74%
E	394	47	53	248	89.34%	88.14%	86.52%	88.74%
BC	397	44	57	244	90.02%	87.44%	86.39%	88.72%
ABDE,ABC	397	44	57	244	90.02%	87.44%	86.39%	88.72%
C	393	48	52	249	89.12%	88.31%	86.52%	88.71%
AD,AC	393	48	52	249	89.12%	88.31%	86.52%	88.71%
BCE,ABCE,ABE	397	44	58	243	90.02%	87.25%	86.25%	88.62%
CE	393	48	53	248	89.12%	88.12%	86.39%	88.61%
A	393	48	53	248	89.12%	88.12%	86.39%	88.61%
ACE	393	48	53	248	89.12%	88.12%	86.39%	88.61%
ADE	392	49	52	249	88.89%	88.29%	86.39%	88.59%
BD	400	41	63	238	90.70%	86.39%	85.98%	88.50%
AE	392	49	53	248	88.89%	88.09%	86.25%	88.49%
BCD,BCDE	399	42	63	238	90.48%	86.36%	85.85%	88.37%
D,CD,CDE	395	46	58	243	89.57%	87.20%	85.98%	88.37%
BDE	399	42	64	237	90.48%	86.18%	85.71%	88.27%
DE	394	47	59	242	89.34%	86.98%	85.71%	88.14%
CDEF,ACDEF	428	13	111	190	97.05%	79.41%	83.29%	87.35%
ADEF	426	15	110	191	96.60%	79.48%	83.15%	87.21%
CDF,ACDF	427	14	112	189	96.83%	79.22%	83.02%	87.14%
DEF	427	14	113	188	96.83%	79.07%	82.88%	87.05%
DF,ADF	424	17	112	189	96.15%	79.10%	82.61%	86.80%
F,CEF,EF,BF,BCEF,BEF,ABF, ABCEF,ABEF,AF,ACEF,AEF	434	7	148	153	98.41%	74.57%	79.11%	84.85%
CF,CBF,ABCF,ACF	434	7	149	152	98.41%	74.44%	78.98%	84.77%
ABDF, ABDEF	423	18	179	122	95.92%	70.27%	73.45%	81.11%
BDF,BDEF	419	22	179	122	95.01%	70.07%	72.91%	80.65%
BDCF,BCDEF, all features, ABCDF	441	0	301	0	100.00%	59.43%	59.43%	74.56%

Note: TP, FP, FN and TN represents true positive, false positive, false negative and true negative in a standard confusion matrix, respectively. A, B, C, D, E, F represents SCP, lift, conviction, confidence, support, and chi-square, respectively. The table is ordered by F-measure, and combinations with the same results are grouped together

of “Aminoacido-pathies”. Another type of mistake is partial translation such as “症候群”. It was a mistake in our candidate generating system, which did not get the non-Chinese character part of that Chinese translation. Both of the above examples, have very high association measure results, high frequency, and not too far distance. Those results caused the wrongly translated Chinese candidates to be classified into a correct class.

For the case that the correct instance is classified into a wrong class, the errors were usually caused by the search engine. For queried term 268 (Myotubular Myopathy 肌小管病/肌小管病变), our retrieved answer is “肌小管病”, but when Yahoo API checks for co- occurrence the result is zero, the co-occurrence for “肌小管病变” is more than 100. Similar searching results with Google and MSN were obtained. So far, we can only assume that those search engines use a statistically based Chinese word segmentation system, the system highly depends on the corpus that occurs on the Internet.

5. Conclusion

We proposed a new approach for solving English medical OOV terms. To our knowledge it is the first approach that handles Chinese translation with non-Chinese characters. It takes into consideration each input English OOV term and adjusts different candidate generation rules accordingly. Unlike most existing methods for translating English OOV term into Chinese, our candidates are selected by machine learning system with support from different features, rather than a rule based candidate selecting system. The new approach has significant improvements over the existing methods. By testing it with two very difficult medical OOV terms from ICD9- CM and ICD9, we are able to reach up to 89.98% precision compared to the existing method with only 61.61%. The improvements are mainly due to better generated Chinese candidates.

6. Future Works

This new approach suggested a better set of generated Chinese candidates to improve the overall English-Chinese OOV term translation performance. Our future works will be mainly focused on how to generate Chinese candidates at a higher precision.

7. References

[1] Y. Zhang and P. Vines, “Using the web for automated translation extraction in cross- language information retrieval,” In SIGIR’04, pages 162-169, ACM Press, 2004.

[2] Y. Zhang, P. Vines and J. Zobel, “Chinese OOV Translation and Post-translation Query Expansion in Chinese–English Cross-lingual Information Retrieval,” In TALIP’05, pages 57-77, ACM Press, 2005.

[3] P.-J. Cheng, et al., “Translating unknown queries with web corpora for cross-language information retrieval,” In SIGIR’04, pages 146-153, ACM Press, 2004.

[4] F.-H. Peng, F.-F. Feng and A. McCallum, “Chinese segmentation and new word detection using conditional random fields,” In ACL’04, Article No. 562, ACL Press, 2004.

[5] Y. Zhang and P. Vines, “Detection and translation of oov terms prior to query time,” In SIGIR ’04, pages 524-525, ACM Press, 2004.

[6] C. Lu, Y. Xu and S. Geva, “Translation disambiguation in web-based translation extraction for English-Chinese CLIR,” In SAC’07,

pages 819-823, ACM Press, 2007.

[7] J. Qu, et al., "Automatic English to Chinese Translation of Medical Terms using Association Rule Mining with Web Data," In JCSSE'09, pages 336-341, 2009.

[8] Silva, J.F., et al., "Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units," Lecture Notes in Artificial Intelligence 1695, pages 113-132, 1999.

[9] T. Limungkura, T. Theeramunkong and P. Aimmanee, "Extraction of Medical Exper- Affiliation Relatinos from WWW," In JCSSE'09, pages195-199, 2009

[10] D. A. Hull, and G. Grefenstette, "Querying Across Languages: A Dictionary-Based Approach to Multilingual Information," In SIGIR'96, pages 49 – 57, ACM Press, 1996.

[11] W.-H. Lu, L.-F. Chien, and H.-J. Lee, "Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach," In TOIS'04, pages 242-269, ACM Press, 2004.

[12] Y. Zhang, F. Huang and S. Vogel, "Mining Translations of OOV Terms from the Web through Cross-lingual Query Expansion," In NLP-KE'03, pages 669-670, ACM Press, 2005.

[13] H.-H. Chen, C.-H. Yang and Y. Lin, "Learning formulation and transformation rules for multilingual named entities," In ACL'03, pages 1-8, ACL Press, 2003.

[14] K. Viriyayudhakorn, T. Theeramunkong and C. Nattee, "Mining Translation pairs for Thai-English Medical Terms," In KICSS'08, 3rd KICSS, 2008.

[15] H.I. Witten and E. Frank, "Practical machine learning tools and techniques with Java implementations," In SIGMOD' 02, pages76-77, AMV Press, 2002.

[16] D. M. Magerman, B. Beranek and N. Inc, "Statistical decision-tree models for parsing," In ACL'95, pages 276 – 283, ACL Press, 1995.