Title	OOV用語処理に基づく翻訳,情報抽出,クロスランゲージ情報検索に関する研究
Author(s)	区, 建
Citation	
Issue Date	2013-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/11550
Rights	
Description	Supervisor:島津　明, 情報科学研究科, 博士

A study on translation, information extraction and CLIR based on OOV term processing

Japan Advanced Institute of Science and Technology
School of Information Science

Jian Qu

A study on translation, information extraction and CLIR based on OOV term processing

by

JIAN QU

Submitted to
Japan Advanced Institute of Science and Technology
In partial fulfillment of the requirement
for the degree of
Doctor of philosophy

Supervisor: Professor AKIRA SHIMAZU

School of Information Science
Japan Advanced Institute of Science and Technology

September 2013

Acknowledgement

It has been a challenging process to achieve a research based doctor degree, especially in one of the best universities in Japan. Upon my graduation, I wish to express my deepest gratitude to my mother-Lecturer YuanJun Zhang who brought me to this world and has been solely taken care of me since I was at the age of ten, her continues support made me capable of achieve this academic goal. This academic goal would never be possible without the support of my supervisor-Professor Akira Shimazu who offered me a chance to continue my research in one of the best schools in Japan, he inspired and guided me along the way, his kind support during my research made me possible of completing this doctor degree. My vice supervisor-Associate Professor Kiyoaki Shirai and Associate Professor Nguyen Le Minh guided me on conducting many of my researches. My sub-theme supervisor Associate Professor Takaya Yuizono inspired me on some special topic and we were able to get a best paper award. This research goal would never be achieved without the support of my former supervisor Associate Professor Thanaruk Theeramunkong. I would like to express my profound gratitude to Professor Akira Shimazu, Associate Professor Kiyoaki Shirai, Associate Professor Nguyen Le Minh, Associate Professor Takaya Yuizono and Associate Professor Thanaruk Theeramunkong. I would also like to express my profound gratitude to Professor Satoshi Tojo and Professor Fuji Ren for helping us perfecting this thesis. I would like to express my sincerely gratitude to my wife-YeZhuang Lu for her continues support. I would also like to express my sincerely gratitude to my beloved dog-Jam, who has lived with me for 16 years and has passed away in Thailand while I was working on my research in Japan. Finally, I would like to express my profound gratitude to JAIST GRP program, which supported many Thai PHD students for conducting their researches in JAIST.

Abstract

OOV term translation plays an important role in natural language processing. Although many researchers in the past have endeavored to solve 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 English definition 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 English definition ranking method also extracts multilingual context information and monolingual definitions of OOV terms. Moreover, non-existing methods focus on cross language definition retrieval for OOV terms. We propose a novel CLIR for Chinese definition retrieval method for extracting Chinese definitions of OOV terms. Never the less, it has always been so difficult to evaluate the correctness of an OOV term translation and definition without domain specific knowledge and correct references. We propose a novel auto re-evaluation method to evaluate the correctness of OOV translations and definitions.

We tested our methods with both name type and biomedical type OOV terms. We retrieved and processed a total of 743,914 documents (snippets). Our method achieved accuracies of 84.15% for multilingual context information extraction and 75.46% for English definition extraction respectively. Our method also achieved precision of 79.76% and high recall of 99.86% for name type OOV term prediction, and we achieved high precision of 99.93% and recall of 89.21% for biomedical type OOV term prediction. For name type OOV term translation, our method gained little improvements over existing methods with high accuracies of 98.39% and 98.39% in candidate generation and candidate selection respectively. For biomedical type OOV term translation, our method gained much improvements over existing methods, our method of SF+F+W+S+B+P with the base machine learning algorithm Lib-supported vector machine surpasses the existing methods with a recall of 83.05% and precision of 79.72% for OOV translation. Furthermore, our method achieved accuracies of 67.49% for Chinese definition extraction, 85.12% for name type OOV term Chinese definition extraction and 60.00% for biomedical type OOV term Chinese definition extraction. We achieved a precision of 46.99% and a recall of 99.37% for translation auto re-evaluation. We also achieved a precision of 67.90% and a recall of 99.41% for definition auto re-evaluation.

Table of Contents

CHAPTER TITLE **PAGE**

List of Figures

List of Tables

Chapter 1

Introduction

One of the most important inventions of the 20th century was the Internet, which has completely changed the way of how human access information and knowledge. People no longer depend on TVs, radios, newspapers, and books. We can get almost any information from the Internet. However, the use of the Internet is limited by one's own language ability. For example, a person who knows only Thai may only understands the Thai webpages. Although there were many Cross Language Information Retrieval (CLIR) tools proposed, they work well only with simple words or short sentences, such as Google translate. A major problem of the CLIR is the Out Of Vocabulary (OOV) terms, which are typically new terms that cannot be found in dictionaries, such as personal names, place names, new technical terms and translated words etc.

Many researchers in the past have endeavored to solve the OOV term translation problem [13, 46, 85-88]. However, new OOV terms are emerging every day, a perfect method that able to handle different types of OOV terms, especially biomedical type OOV terms is yet to be discovered. Moreover, the translation of an OOV term is just another OOV term in different language, without additional information users cannot understand the meaning of such OOV term. Furthermore, not only extracting translation and context information but also extracting definition of OOV term has become important lately. Moreover, it has always been very difficult to evaluate the correctness of OOV translations and definitions. We propose an OOV term type prediction method, two OOV term translation methods, a multilingual context information extraction method, a monolingual definition extraction method, a cross language definition extraction method and an auto re-evaluation method for OOV terms using rule-based pattern extraction, web-page pattern recognition, supervised machine learning and artificial intelligent algorithms.

In this chapter, we are going to describe the cross language information retrieval, definition extraction and the out of vocabulary problems, our motivation, objectives, scope of our work, and thesis overview, in 1.1, 1.2, 1.3, 1.4, and 1.5 respectively.

1.1 Cross language information retrieval, Out of vocabulary translation, and definition extraction

Cross language information retrieval aims to retrieve target language documents from source language queries. CLIR employs multilingual dictionaries to achieve such task. OOV term is a large problem for dictionary based CLIR, because multilingual dictionaries have a low inclusion rate of daily emerging OOV terms. Solving OOV term problem would improve the performance of CLIR.

Out of vocabulary terms are typically new terms from current affairs, such as personal names, place names, newly technical terms and translated words. Compound words can be considered as another type of OOV terms. A compound word is a word that composes of multiple sub-words that constitute a new meaning. This compound word may not relate to the meanings of its sub-words, consequently direct translation of each individual word would result in miss translation. For example, the word “cross straits” means China and Taiwan relationship that is totally unrelated to its component words.

OOV terms can be classified into two groups, 1) the name type OOV terms such as personal names, brand names, location names, and compound words, etc. And 2) the technical type OOV terms such as newly technological terms and biomedical terms, etc. Name type are usually translated by transliteration, and the Chinese translation usually includes only Chinese characters. Technical type OOV terms are usually translated by a combination of meaning and transliteration.

Definition extraction is a new field in natural language processing. Definition extraction is a subtask of terminology extraction. It is important for natural language processing. It is useful for question answering, automatic creation of glossaries, learning lexical semantic relations and automatic construction of ontologies.

1.2 Motivation

Out of vocabulary term has always been a problem for natural language processing, especially for Cross language information retrieval. Given an OOV term in source language, OOV term translation extraction aims to find the correct translation in target language. CLIR can be improved when OOV term is correctly translated.

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. 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 biomedical terms end with the term "症候群". Another fundamental problem for existing approaches is the translations obtained in other languages are also OOV terms. These translations provided little information to users 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. Furthermore, not only extracting translation and context information but also extracting definition of OOV term has become important lately. Moreover, non-existing methods focus on cross language definition extraction. Never the less, it has always been so difficult to evaluate the correctness of the OOV translation and definition without correct references.

1.3 Objectives

We propose to translate name type OOV terms and biomedical type OOV terms using different approaches. Thus, we will predict the types of OOV terms before translating them. We propose a novel English ranking approach to automatically predict the types of OOV terms. Then a novel adaptive rules approach together with supervised machine learning is used for finding translations for biomedical type OOV terms, and we propose a novel statistical filter together with existing ranking list approach for finding translations for name type OOV terms. Furthermore our novel English ranking approach also extracts multilingual context information and monolingual definitions for both name type and biomedical type OOV terms. Moreover, a novel CLIR for Chinese definition ranking algorithm is proposed to do a cross language definition retrieval of English OOV terms. Furthermore, we propose a novel auto re-evaluation algorithm to evaluate the correctness of OOV translations and definitions. For example, "Mae West", our method 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), we would also extract its definition in English (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...), definition in Chinese (梅·蕙絲(英语：Mae West，1893年8月7日－1980年11月22日)，演員、劇作家、螢幕編劇、也是美國眾所週知的性感偶像。梅·蕙絲最為人所熟知的是其黃色的雙關語，在到好萊塢為電影產業寫劇本、做諧星、演員之前，她就在紐約的劇場舞台上以綜藝...), Chinese definition auto evaluation score of 0.6, and Chinese translation auto evaluation score of 1.

1.4 Our contribution

To the best of our knowledge, our proposed method is the first available system that makes it possible to distinguish different types of OOV terms before translation, and we proposed different methods to find translations for each type of OOV terms. Our idea may contribute to many machine translation tasks, which may offer a new idea that grouping potentially different terms into different groups prior to translation. We proposed a novel self-modifying method for solving hybrid translations and we were able to handle both name type and biomedical type OOV terms. Our self-modifying method may help the process of pattern matching in machine translation, which may help related researchers to find a way of auto pattern construction and self-modifying patterns. We tested 24 combinations of machine learning methods on a data set of 157,075 candidates with 25 features, we proposed a novel statistical filter that enabled many expensive machine learning methods possible for large data experiments. Related researchers may use our statistical filter for their machine learning tasks. Furthermore, we proposed novel methods for extracting multilingual context information and definitions for OOV terms. Our multilingual context information extraction method is a general approach, thus can be applied to different languages. Our definition extraction method utilized many web features, thus related researches can utilize these features and improve their work. Moreover, our method is the first available system that does a CLIR of Chinese definition extraction for OOV terms. It helps user to understand the meaning and concept of OOV terms. We utilized many web features and made some discoveries in CLIR for Chinese definition extraction. We believe related researchers may use our discoveries and improve their works. Finally we proposed a novel auto re-evaluation method for Chinese definition and Chinese translations. This work offers a new idea for auto re-evaluation, related works may use such idea in their researches. The auto re-evaluation is a general method, thus can be applied to different languages. The overall of our works greatly contribute to machine translation, term translation, information retrieval, information extraction, cross language information extraction and machine learning.

1.5 Scope of our work

This thesis focuses on name type and biomedical type OOV term translation from English to Chinese. Moreover, this thesis also investigates the methods for extracting multilingual context information and monolingual definition for OOV terms. Furthermore, this thesis investigates cross language information retrieval of definition extraction for OOV terms. Finally, this thesis investigates the methods for generating confidences scores for predicting correctness of OOV translations and Chinese definitions. This thesis looks in detail of the existing English OOV term to Chinese translation approaches and finds ways to develop new methods to improve the quality of the results in terms of accuracy. This thesis also briefly investigates the existing term definition extraction methods, and develops novel methods for extracting multilingual context information and monolingual definitions for OOV terms. This thesis also briefly investigates the cross language information retrieval for OOV terms, and develops new methods for cross language information retrieval of definition extraction for OOV terms.

1.6 Thesis overview

The overall thesis is outlined in the following chapters. The related works are presented in Chapter 2, which is a summary of extensive literature review. In Chapter 3, we present our proposed approaches. In Chapter 4, we present the experimental results of our proposed approaches. In Chapter 5, we present the detailed discussions of those experimental results. Finally, we conclude the thesis in Chapter 6.

Chapter 2

Related works and background

In this Chapter, we will firstly introduce knowledge regarding Chinese language, then we will review the observation, existing methods and problems of OOV term translation, OOV term definition retrieval and OOV terms in CLIR, in 2.1, 2.2, 2.3, 2.4, 2.5, and 2.6 respectively.

2.1 Introduction to Chinese language

Chinese language is a language spoken by more than one billion people over the world, is a language that constitutes the second largest number of speakers. The old Chinese Qin Empire regulated the Chinese language or Standard Mandarin for more than 2000 years ago. Standard Mandarin is the official language of the People's Republic of China (PRC) and the Taiwan. Although Chinese and Taiwanese all speaks the same Standard Mandarin, they have different writing systems.

For the Standard Mandarin speaking language, there are two different writing systems, which are Traditional system and Simplified system. The traditional system is usually very difficult to write, as there are many strokes for writing. The traditional system comes with the Chinese characters based phonic system for pronunciation purposes. This system is used in Taiwan, Hong Kong, and Macau as the official written system and some other Chinese speaking communities including Thailand.

The simplified Chinese system is modified from the traditional system to reduce the number of strokes for each Chinese character in order to write them more easily. The simplified system uses a Romanic phonic system for pronunciation purposes. This system is used in China and Singapore as the official written system.

The traditional writing system writes the Chinese characters-Hanzi in vertical columns from top to bottom whereas the simplified writing system writes the Chinese character in horizontal rows from left to right. Nowadays, texts both traditional and simplified writing system are usually arranged in horizontal rows from left to right.

Unlike English, Chinese language has no word boundary between Chinese words. In addition, almost all Chinese characters have a meaning and thus they are treated as a word. According to a comprehensive Chinese dictionary “Hanyu Da Cidian,” it contains 23,000 Chinese characters and these characters can form more than 370,000 Chinese words.[25, 42, 74]

The largest challenge of Chinese language for NLP is word segmentation, if a perfect word segmentation exists for Chinese language, many problems including OOV term translation problems would be solved without further research. Another important problem is that many Chinese language characters share same sounds. Moreover, Chinese OOV translations are not unified, especially for terms that using transliteration method. Thus there may be few correct Chinese translations for an English OOV term available on the Internet. Each Chinese translation many use characters in similar sounds, but they are different in writing, for example, a name type OOV term "Benz" with its un-unified Chinese translations"奔驰" and "賓士".

2.2 Observation of OOV term translation

The out of vocabulary are words arise from daily actives, such as personal names, location names, brands, newly technology terms and translated words. Given a referenced dictionary, any words that cannot be found in the dictionary are called out of vocabulary terms. OOV terms can be compound words in which each component is in a dictionary but the combination of those words gives a new meaning. Such OOV terms are idioms or proper names for example, “cross straits”.

From its definition, whether or not a term is an OOV term depends mainly on the dictionary that it is used for reference. One word may be an OOV term for one dictionary, but may not be an OOV term for another dictionary. Because of the dictionary dependence, OOV words are often a problem for automatic spell checkers, voice recognition software, cross language translation software and cross language information retrieval.

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. Name type OOV terms are usually translated by transliteration. Some translations of brand name OOV term use a combination of semantic translation and transliteration. Technical type, especially biomedical type OOV terms are usually translated by a combination of semantic translation and transliteration. However, some translations of biomedical type OOV terms are hybrid translations. Hybrid translations use part target language and part source language, for example, English OOV terms and their Chinese translations: “DiGeorge's syndrome” (DiGeorge's 症候群) and “α1-antitrypsin deficiency” (α1-抗胰蛋白酶缺乏症).

2.3 Existing approaches for OOV term translation

Generally speaking, OOV terms can be translated by six methods, they are: 1) human translation with field of specified knowledge; 2) automatic translation using domain-specific bilingual dictionaries [56]; 3) automatic translation using transliteration [38]; 4) automatic translation using both bilingual dictionaries and transliteration; 5) automatic mining from parallel corpus or comparable corpus [90]; 6) and automatic mining from the Internet [46]. Human translation has always been the best method for high quality translations, however the cost are extremely high. Many traditional OOV term translation methods using bilingual dictionaries, IBM 1-3 model is a typical lexicon mapping method, which utilizes bilingual dictionaries to translate OOV terms [81]. Although lexicon mapping offers low cost and high accuracy but it results in low recall. Transliteration method employs phonetic mapping algorithm for OOV term translation. It is very successful on languages that share a similar phonetic system, such as English and French [2], but not very successful on languages that have very different phonetic system such as English and

Chinese. A combination of lexicon mapping and transliteration method offers the advantage of low cost, high accuracy, and good recall. However the translations generated by such method are usually different from human translations and many not always suitable for CLIR. Recent researches show automatic mining takes the advantage of both low cost and high quality translations. Automatic mining utilizes pattern matching or statistical rules for mining the human translations from parallel corpus or comparable corpus. However, bilingual corpuses are not always up to date, automatic mining from the Internet solves such problem. Most OOV terms have their correspondent human translations on the Internet [13, 22, 84]. The translations mined from the Internet are usually high quality and require low human cost. Many existing OOV term translation approaches adapt either automatic mining or web mining[31, 65, 72].

Automatic mining can be developed into four steps, they are:

1) data collection aims to collect the documents containing the possible translations of the OOV terms from the Internet, parallel corpus or comparable corpus;

2) Translation candidate extraction aims to find the boundaries of the translations in the documents;

3) Translation candidate filtering aims to reduce the amount of wrong translation candidates;

4) Translation candidate selection aims to choose the correct translations from the extracted translations. In the following subsection, we review those four steps of automatic mining.

2.3.1 Data collection

Many multi-language documents are available on the Internet; most of them come from the technical research papers, organizational or governmental web sites. Zhang and Vines stated if English OOV terms occur in Chinese webpages, the Chinese translations for these English OOV terms are usually nearby [86]. Cheng *et al.* observed that if a Chinese term occurs in an English webpage, its translation usually exists in the same page as well [14]. A number of existing approaches have presented techniques for retrieving the translation pairs from multi-language documents. Those methods can be summarized as follows.

2.3.1.1 The parallel corpus based approach

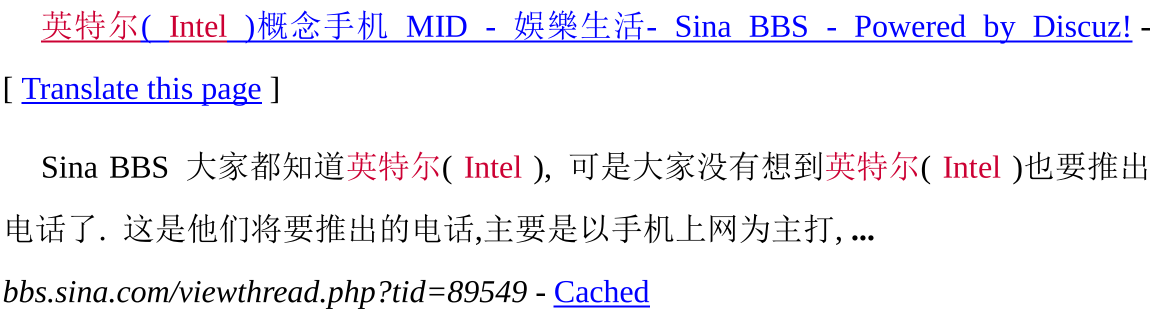
Given an OOV term in the source language and a known text containing the OOV term, this approach finds its translation of the OOV term in the target language by utilizing an available translated text of the known document in the target language. Such a pair of documents is called parallel texts. This concept is derived from the concept of parallel multi-language texts, which are used widely in the field of cross language information retrieval [24, 40, 51, 61, 78].

2.3.1.2 The anchor text based approach

Proposed by Lu *et al.* in 2004 to solve the translation problem of web queries, this approach uses the web anchor texts and its link structure to extract translations of web queries. The anchor texts are set of hyperlinks that contain the URL, the title, and the summary to the original website. Collections of anchor texts from different languages are referring together to the same page. When the OOV terms are inputted in the search engine, the returned anchor texts can link to a number of webpages in a target language that possibly contain the translation of our OOV term query [46].

2.3.1.3 Web query based approaches

This approach is a widely used method to mine an OOV term in the source language and its translation in the target language. It is done by querying a source OOV term to a search engine and specifying results to be in the target language. The returning search results are called “snippets.” The snippets contain brief but useful information about title, URL, and the sample of output containing the OOV terms. In the close location to the OOV term appeared in the retrieved sample text, the candidates of OOV term translation may be found. A snippet of OOV term query “Intel” is shown in Figure 2.1 [14, 46, 86].



英特尔(Intel)概念手机 MID - 娱樂生活- Sina BBS - Powered by Discuz! - [Translate this page]

Sina BBS 大家都知道英特尔(Intel), 可是大家没有想到英特尔(Intel)也要推出电话了. 这是他们将要推出的电话,主要是以手机上网为主打, **...**

bbs.sina.com/viewthread.php?tid=89549 - Cached

Figure 2.1: An example of snippet from query “Intel”

2.3.1.4 Query extending approaches

The web query based approach may have a big drawback of not always getting the translation due to not enough information. Proposed by Zhang *et al.* in 2005, this query extension approach expands the original OOV terms in the source language by adding some hints in the target language before querying them to the search engine. This way would increase the possibility of getting the texts containing the OOV terms [83].

2.3.2 Translation candidate extraction

Candidate generation or candidate extraction consists in determining the boundaries of the Chinese translation in the snippets. Given texts that contain possible target translation of the source OOV term, in this step we try to generate the target translation candidates for the source OOV term. Because in those web retrieved texts, they usually contain the correct translation term. But automatic generating the translation candidates is always difficult, especially for certain languages like Thai, Japanese, including Chinese which do not have a clear word boundaries. Existing approaches suggest some text preprocessing such like word segmentation is not necessary because the translation for the OOV term is not in the dictionary. Word segmentation tools are based on dictionaries.

Existing methods of extracting the translation from the snippets can be classified into three groups; the first group uses the brute force method to generate all possible combinations of the Chinese characters in the snippets [13, 45, 84-86]. This method is inspired by Zhang and Vines [86], they focus on name type OOV terms, such as personal names, brand names, and organization names.

The second candidate generation method tries to find patterns that identify the translations in the snippets [73]. For example, the text appearing in parentheses is usually the correct translation.

The third group uses word/sub-word, and lexical mapping with dictionaries in order to solve the OOV term problem. The details of each method are provided in the following subsections.

2.3.2.1 Considering all substrings

One of the most accurate and widely used approaches for generating Chinese translation candidates for English OOV terms was proposed by Zhang and Vines [86]. They used a candidate length & co-occurrence method and achieved up to 80% accuracy to translate English name type OOV terms such as personal names, company names and brand names. Given a snippet containing the source language OOV term in target language, this method first removes all punctuations and counts the numbers of OOV term. When a source language OOV term is found, they collect up to thirty target language characters in front of and behind the source language OOV term. Since segmentation system is not used here. They have to consider all substrings from both front and back part. For example, a string containing Chinese characters and the English OOV term: “可是大家没有

想到英特尔(Intel)也要推出电话了”. This method generates all substring of the front part: “可是大家没有想到英特尔” and back part: “也要推出电话了”, for example, the substring of the back part are: 也, 要, 推,出, 电, 话, 了, 也要, 也要推, 也要推出, 也要推出电, 也要推出电话了, 要推出电话了, 也推出电话了, 出电话了, 电话了, and 话了. They collect the frequency and the candidate length for each substring for later use [14, 45, 85]. This method yields a very high recall, but also obtains a low precision due to large amount of Chinese translations, thus it becomes difficult to select the correct translations. In order to increase the precision and filter the amount of translations, Cheng *et al.* suggested the Symmetrical Conditional Probability and Context Dependency (SCPCD), Lu *et al.* used the Symmetrical Conditional Probability (SCP), and Zhang *et al.* used local maximum [13, 45, 88].

2.3.2.2 Pattern translation candidates generating system

Viryayudhakorn argues that the above method generates too many incorrect target language translation candidates. They suggest that most OOV translations are usually enclosed in a parenthesis and placed immediately after the source OOV term. They generated few parenthesis patterns according to document observation. Given a snippet containing the source language OOV term in target language, this method scans for parenthesis patterns near the OOV term. They extract all characters within the parenthesis that matches to their patterns [73]. However, as texts available on the website are usually disorganized, they appear in different forms, hence this method usually yields a low recall, because some correct translations do not appear in the parentheses. In addition, this method produces very few substrings of the correct translation, which result in a high precision. Some examples of the snippets retrieved from the web that contain English biomedical OOV terms and their Chinese translations are shown in Table 2.1. According to these examples, we can conclude that there is no standard pattern for Chinese biomedical translation on the Internet.

Table 2.1: Some examples of snippets of English biomedical OOV terms

	Conditions	Snippets	Methods	Eoov	Translation
Eoov before Translation					
	Eoov+noise+ Traslation	6, 1, Hypermethioninemia, 高甲硫胺酸血症, 270.4. 7, 2, Cystinosis, 胱胺酸症, 270.0. 8…	**2,3**	Hypermethioninemia	高甲硫胺酸血症
	Eoov+noise+ Chinese+Traslation	Urea Cycle Disorders ●. 一、什麼是尿素循環障礙. 生化機轉 哺乳類動物蛋白質代謝後的主要產物為尿素,攝入愈多蛋白質食物,產生之尿素也愈多。	**1,3**	Urea Cycle Disorders	尿素循環障礙
	Eoov+noise+ English+Traslation	755.59. 45. DiGeorge's syndrome.ICD9-CM. DiGeorge's 症候群.	**1**	DiGeorge's syndrome	DiGeorge's 症候群
	…	**…**		**…**	**…**
Translation before Eoov					
	「Translation +（Eoov）」	【大紀元 4 月 2 日訊】（自由時報記者陳賢義／台東報導）台東縣海端鄉一對接續罹患罕見疾病「異戊酸血症（lsovaleric acidemia）」的余姓小姊妹，因母親去年病逝，照護 ...	**2**	lsovaleric acidemia	異戊酸血症
	Translation+Chinese+(Eoov)	非酮性高甘胺酸血症診斷與治療(Nonketotic hyperglycinemia).	**1**	Nonketotic hyperglycinemia	非酮性高甘胺酸血症
	Translation+noise+Eoov	0312，原發性肉鹼缺乏症, Carnitine deficiency syndrome, primary, 0326，丙酮酸鹽脫氫酶缺乏症, Pyruvate dehydrogenase deficiency …	**2,3**	Carnitine deficiency syndrome	原發性肉鹼缺乏症
	…	**…**		**…**	**…**

Note
Eoov: English OOV term
Methods: which existing method described in 2.3.2 can be applied to detect the Chinese translation
Dark(green) highlights: Chinese translation
Light(yellow) highlights: English OOV term

2.3.2.3 Word/sub-word and lexical mapping system

Zhang and Sumita have suggested the employment of the lexical mapping for the OOV term, they split longer OOV terms into smaller pieces, and translate the smaller pieces term by term or character by character according to the dictionary. This method usually yields extremely high precision for OOV terms that are made of known sub-words, for example, “at 3:15” (三点十五分) or “national aeronautics and space administration”(国家航空航天局) [81], and may gain a low recall if OOV terms are made of unknown sub-words. The performance of the method is also very much depending on the dictionaries consulted.

2.3.3 Translation candidate filtering

There are always too many possible translation candidates to process forward, so in this step we present several existing techniques to filter possible translation candidates. There are two major approaches, which are co-occurrence statistics and mutual information, where the co-occurrence statistics is the most widely used approach.

2.3.3.1 Co-occurrence statistics

Given a list of tremendously many possible translation candidates generated by all substrings of the retrieved snippets, a ranking list function is used for ranking the substrings based on the likelihood of being the correct translation. The ranking list function R_i can be described as in the following formula [86]:

$$R_i = \alpha \times \frac{|C_i|}{\max_j |C_j|} + (1-\alpha) \times \frac{f(Ci)}{f(Eoov)} \tag{2.1}$$

where $\max_j |C_j|$ is the maximum length of the target language substring found. $|C_i|$ is the length of each possible target language translation candidate. $f(Ci)$ is the frequency of the target language translation candidate and $f(Eoov)$ is the frequency of the source language OOV term. According to Zhang and Vines’s experiments, α is a value that is proved to be robust across their experiments, they suggested that α = 0.25 provides the best combination of frequency and length. If α is big, the ranking function gives high rank to longer candidates; if α is too small, the ranking function gives high rank to candidates that are more frequent.

According to the ranking list function, the candidate that appears frequently (large $f(Ci)$) and has long length (large$|C_i|$) is getting higher rank than the candidates that rarely appear and has short length. Because the method is built specifically for OOV term translations, these translations are usually longer than normal terms.

This approach is widely used as it can quantify the major features: frequency and length that affect the OOV term translation, the major disadvantage of this approach is that only two features of the possible translation candidates are focused, therefore it usually yields results with low overall recall and precision.

2.3.3.2 Mutual information

Simply selecting the longest translation candidates and the most frequent translation candidates do not always give the correct outcomes. Several approaches try to improve this problem by ranking all the translation candidates with some probability model.

Given a list of excessively many possible translation candidates generated by all substrings of the retrieved snippets, we need to filter them by mutually consider the co-occurrence frequency of each character in the possible translation candidates. Mutual information is usually applied to quantify the distance between the joint distributions of two terms. Since the list of possible translation candidates contains all substrings, some of them maybe a term and some of them maybe a sentence. Lu *et al.* suggested a few approaches for determining the candidate to be either word or sentence, two methods using mutual information to quantify the likelihood of the word are explained below [35, 45].

1. Symmetrical Conditional Probability (SCP)

Given a list of possible translation candidates from all substrings of the retrieved snippets, we can rank each translation candidates by using Symmetrical Conditional Probability (SCP). SCP score ranges from 0 to 1. For each substring in the possible translation candidate, SCP calculates the frequencies of these substrings in the corpus [69]. Then SCP compares these frequencies with the frequency of the possible translation candidate. The SCP score is close to 1 when the substring of the possible translation candidate occurs less often in the corpus than it occurs only within the translation candidate itself. If a string has a high SCP score, the string is likely to be a word. SCP can be determined in the formula below [45, 50].

$$SCP(c_1 \dots c_n) = \frac{(n-1)f(c_1 \dots c_n)^2}{\sum_{i=1}^{i=n-1} f(c_1 \dots c_i) f(c_{i+1} \dots c_n)} \tag{2.2}$$

SCP takes string$(c_1 \dots c_n)$ as the input where$(c_1 \dots c_n)$ is a string in target language which is the possible translation candidates, n is the number of characters in the string. For each input we separate it into two substring, the front substring $(c_1 \dots c_n)$ and back substring$(c_{i+1} \dots c_n)$for all possible patterns of the input. $f(c_1 \dots c_n)$ is the frequency of that string, and $f(c_1 \dots c_n)$ and $f(c_{i+1} \dots c_n)$ are the frequency of front substring and back substring.

For example, a Chinese string containing English OOV term: “大家都知道英特尔 (Intel), 可是大家没有想到英特尔 (Intel) 也要推出电话了.” One possible Chinese translation candidates from this string is 英特尔. SCP tries to determine if 英特尔 is a word or a sentence by comparing the frequencies of each substring of 英特尔 to the frequencies of 英特尔. An example of calculation is shown in Table 2.2:

Table 2.2: An example of SCP calculation

Possible translation candidates	Frequency in the corpus	Substrings of the candidate	Frequencies in the corpus
英特尔	2	英	2
		特	2
		尔	2
		英特	2
		特尔	2

$$SCP(\text{英特尔}) = \frac{(3-1)(2)^2}{(2)(2)+(2)(2)}$$

$$SCP(\text{英特尔}) = 1$$

SCP is used for filtering the possible translation candidates, the filter process starts with the strings with the longest length, since they are usually more likely to be a sentence than a word.

1. If the longest string with n characters resulted a higher SCP score than its n-1 substring, that longest string is recorded as a translation candidate, and all of its substring's frequencies are deducted by this longest string's frequency.

2. If the longest string with n characters result a lower SCP score than its n-1 substring, this longest string is removed, and we take the n-1 substring as the new longest string and repeat the above test.

SCP can help to differentiate the type of string whether or not it is a sentence or a word. However, if the corpus is very large, or the translation of the OOV term contains frequently used substrings, SCP is not being able offer a high score for the correct translation candidates.

2. Symmetrical Conditional Probability & Context Dependency (SCPCD)

Given a list of possible translation candidates from all substrings of the retrieved snippets, we can filter each translation candidates by using Symmetrical Conditional Probability & Context Dependency (SCPCD) [14] which not only considers the possibility of a candidate to be a word but also considers the sentence containing this candidate is a word or sentence. In SCPCD, symmetrical conditional probability (SCP) is used to measure how likely the candidate is a truly word term to a certain degree [68] and context dependency (CD) [15, 28] looks at the sentence containing this candidates in the corpus and checks whether or not that sentence is a word or sentence. CD also ranges from 0 to 1. SCPCD can be calculated as follows:

$$SCPCD(c_1 \dots c_n) = SCP(c_1 \dots c_n) \times CD(c_1 \dots c_n) \qquad (2.3)$$

where the calculation of SCP and CD are

$$SCP(c_1 \dots c_n) = \frac{(n-1)f(c_1 \dots c_n)^2}{\sum_{i=1}^{i=n-1} f(c_1 \dots c_i)f(c_{i+1} \dots c_n)} \qquad (2.4)$$

$$CD(c_1 \dots c_n) = \frac{LC(c_1 \dots c_n)RC(c_1 \dots c_n)}{f(c_1 \dots c_n)^2} \qquad (2.5)$$

CD takes string $(c_1 \dots c_n)$ which is the target language translation candidate as an input, where $LC(c_1 \dots c_n)$ and $RC(c_1 \dots c_n)$ is the frequency of left and right adjacent target language characters in the corpus. The adjacent characters are the characters next to the target language translation candidate. If the adjacent characters of a target language translation candidate are always same in the corpus that means this target language translation candidate is a substring of another word. If the adjacent characters of a target language translation candidate are different, $LC(c_1 \dots c_n)$ or $RC(c_1 \dots c_n)$ is equal to the frequency of $(c_1 \dots c_n)$.

SCPCD can be described as follow

$$SCPCD(c_1 \dots c_n) = \frac{LC(c_1 \dots c_n)RC(c_1 \dots c_n)}{\frac{1}{n-1}\sum_{i=1}^{i=n-1} f(c_1 \dots c_i)f(c_{i+1} \dots c_n)} \qquad (2.6)$$

where $(c_1 \dots c_n)$ is target a translation candidate, n is the number of characters in the target translation candidate, $f(c_1 \dots c_n)$is the frequency of the target translation candidate, and $f(c_1 \ldots c_n)$ and $f(c_{i+1} \ldots c_n)$ is the frequency of the front part and the back part of the target translation candidate. $LC(c_1 \ldots c_n)$ and $RC(c_1 \ldots c_n)$ is the number of left and right target language characters in the corpus.

For example, a Chinese string containing English OOV term: “大家都知道英特尔 (Intel), 可是大家没有想到英特尔 (Intel) 也要推出电话了.” One possible Chinese translation candidates from this string is 英特尔, this example is shown in Table 2.3.

Table 2.3: An example of SCPCD calculation

Possible translation candidates	Frequency in the corpus	left and right adjacent target language characters	Frequencies in the corpus
英特尔	2	道	1
		到	1

$$SCPCD(\text{英特尔}) = \frac{(2)(2)}{\frac{1}{3-1}((2)(2)+(2)(2))}$$

$$SCPCD(\text{英特尔}) = 1$$

2.3.4 Translation candidate selection

The final process is to automatically select the correct translation for the OOV term. While some closely related languages such as English and French share a very similar grammar, character and sound, the translation selection process can use a transliteration approach. When focus on no closely related languages such as English and Chinese the transliteration approach fails. Existing approaches for candidate selection can be classified into two groups; The first group uses statistical information such as ranking, mutual information, and association rules. The second group utilizes machine learning technologies. We explain some widely used existing methods as follows.

2.3.4.1 Co-occurrence statistics & Longest Common Substring (LCS) approach

Given a list of possible translation candidates generated and filtered from the above steps, we can select the correct translation candidates by using co-occurrence statistics and LCS approach. To select the correct translation candidates Zhang & Vines suggested there are two steps need to be done [86]. First step, the candidate(s) that come with the longest length are chosen. If there is more than one candidate with the same length, the one that has a higher frequency is chosen, but if both length and frequency of two or more candidates are same, all of those will be selected. Second step, the translation candidate(s) that has the highest frequency is chosen. If however there is more than one candidate with the same frequency but if both frequency and length of two or more candidates are same, all of those will be selected. LCS is used to select the candidates that have the longest length with the frequency more than 1, LCS substring is calculated by LCS suffix and those formulas are show below [21, 44, 67]:

$$LCSuff\big((c_1 \dots c_p),(d_1 \dots d_q)\big) = \begin{cases} LCSuff\big((c_1 \dots c_{p-1}),(d_1 \dots d_{q-1})\big) + 1 & if\ c[p] = d[q] \\ 0 & otherwise \end{cases} \tag{2.7}$$

Longest common suffix takes two strings $\big((c_1 \dots c_p),(d_1 \dots d_q)\big)$ as input, where $(c_1 \dots c_p)$ is a string contains p characters; $(d_1 \dots d_q)$ is another string contains q characters, if one character of string $(c_1 \dots c_p)$ is same as another character in string $(d_1 \dots d_q)$, the $LCSuff$ increase one, otherwise, it remains 0 [7, 32].

$$LCSubstr\big((c_1 \dots c_p),(d_1 \dots d_q)\big) = \max_{1\le i\le m, 1\le j\le n} LCSuff\big((c_1 \dots c_i),(d_1 \dots d_j)\big) \quad (2.8)$$

For example, two Chinese strings: “知道英特尔” and “英特尔知道**”.** In order to calculate LCS for these strings, we need to calculate the *LCSuff* first. *LCSuff* checks if first string’s characters are same with second string’s characters [70]. Some examples of *LCSuff* are shown below.

$$LCSuff(\text{知}, \text{英}) = 0$$

$$LCSuff(\text{知道英}, \text{英}) = LCSuff(\text{知道}, \text{""}) + LCSuff(\text{英}, \text{英})$$

= 0 + 1

= 1

$$LCSuff(\text{知道英特尔}, \text{英特尔}) = LCSuff(\text{知道英特}, \text{英特}) + 1$$

$$= \ LCSuff(\text{知道英}, \text{英}) + 1 + 1$$

$$= LCSuff(\text{知道}, \text{""}) + 1 + 1 + 1$$

= 0+1+1+1

=3

The *LCSuff* find that character 知 and 英 is not the same, so it resulted 0. The *LCSuff* find that character 英 and 英 is the same, so it resulted 1. The *LCSuff* calculations are showed in Table 2.4.

Table 2.4 is an example of LCS calculation matrix, where “知道英特尔” is string A and “英特尔知道**”** is string B. The longest common substring of string A and sting B is the max value in the matrix, which is the value of *LCSuff* (知道英特尔, 英特尔). The *LCS* is 英特尔 with a length of 3 characters.

Table 2.4: An example of LCS calculation

String A / String B	知	道	英	特	尔
英	0	0	1	0	0
特	0	0	0	2	0
尔	0	0	0	0	3
知	1	0	0	0	0
道	0	2	0	0	0

The Co-occurrence Statistics & longest common substring (LCS) approach follows a clear logical and selects the longer candidates as the translation candidates. However, sometimes these longer candidates may be a sentence; the method try to fix this by also selects the most frequent candidates.

A disadvantage of this method is that the method can still select more than one translation in some cases.

2.3.4.2 Total correlation approaches

Given a list of possible translation candidates, we can also select the correct translation candidate by finding the possibility of how often each character in the translation candidates occur together in the corpus. The total correlation defines the level of the dependency of each candidate in the corpus. Lu *et al.* suggested the more independent of the candidates, the higher the chance to be the correct translation candidate. A formula for finding total correlation of each character in the possible translation candidates is shown below [45]:

$$Correlation(c_1 \dots c_n) = log_2 \frac{N^{n-1} f(c_1 \dots c_n)}{f(c_1)f(c_2)f(c_3) \dots f(c_n)} \quad (2.9)$$

where $c_i (i = 1 \dots n)$ are target characters in the translation candidates of source OOV term, $f(c_i)$ is the frequency that the c_i appears in the retrieved snippets, $f(c_1 \dots c_n)$ is the frequency that all target characters appear together in the retrieved snippets and N is the size of the retrieved snippets.

For example, a possible Chinese translation 英特尔 for the English OOV term Intel. This example is shown in Table 2.5.

Table 2.5: An example of total correlation calculation

Possible translation	Frequency in the corpus	Substrings of possible translation	Frequencies in the corpus
英特尔	20	英	26
		特	23
		尔	20

Number of the snippets is 100

$$Correlation(英特尔) = log_2 \frac{100^{3-1}(20)}{(26)(23)(20)}$$

$$= 1.223$$

2.3.4.3 The Chi-square method

We can also get the correct translation by Chi-square test, which checks the relationship between source language OOV and the target language translation candidates. Chi-square tests a list of possible translation candidates with their source OOV term. A correlation relationship between a source language OOV term and its target language translation candidate can be measured by this method. The Chi-square tests the hypothesis that source OOV term and its possible translation candidates do not co-occur together[52]. Cheng *et al.* applied the Chi-square test for selection the correct translation candidates with in two steps. First we need to get the number of returned pages obtained by querying the OOV term, translation candidates, or both to a search engine. Second, we can calculate the Chi-square score based on those numbers. We explain the details of those two steps below.

For the first step, each source language OOV term may have more than one translation candidates. For each translation candidate associated, we submit the source OOV term, translation candidate, source OOV term together with the translation candidate to the search engine and record the number of pages returned by those queries. We also need to get the total number of webpages in the world, and the number of pages that does not contain either OOV term nor translation candidates, because the total number shows the ratio between the OOV term and Chinese translations.

For the second step, we need to use the gathered numbers in the first step to compute the Chi-square score. Cheng *et al*. modified the conventional Chi-square to the following formula to suit the web retrieved number of returned pages [12, 14].

$$\chi^2(Eoov, (c_1 \dots c_n)) = \frac{N \times (a \times d - b \times c)^2}{(a + b) \times (a + c) \times (b + d) \times (c + d)} \tag{2.10}$$

where the meaning of each variable is explained as follows

Variable	**Meaning**
$Eoov$	source OOV term
$(c_1 \dots c_n)$	translation candidates
a	$S(Eoov \land (c_1 \dots c_n))$

b $S(Eoov\wedge\neg(c_1 \dots c_n))$

c $S((c_1 \dots c_n)\wedge\neg Eoov)$

d $S(\neg Eoov\neg(c_1 \dots c_n))$

S Is a function that takes a query as the input and returns the number of WebPages containing that query

While the existing candidate selection methods such as mutual information and length co-occurrence analysis tend to use the co-occurrence frequency to be the most important feature. Chi-square test improves much on this co-occurrence frequency by checking not just only co-occurrence but also independency [79]. Chi-square test how often that OOV term exists in the web site without its translation candidates, and how often that translation candidates exist without the OOV term.

For example, a possible Chinese translation 英特尔 of the English OOV term Intel. Their chi-square score is calculated as follows:

Table 2.6: an example of chi-square calculation

$Eoov$	$(c_1 \dots c_n)$	A	B	C	D
Intel	英特尔	40	90	50	10

Number of the snippets is 100

$$x^2(Intel, \text{英特尔}) = \frac{100 \times (40 \times 10 - 90 \times 50)^2}{(40 + 90) \times (40 + 50) \times (90 + 10) \times (50 + 10)}$$

= 23.9458

2.3.4.4 Association measures

Association measures can find the relationship between items in a data set. Follows are introduction and processes of association rule mining for finding relationships between items [82].

Association measures discover the relations of item A and item B in a data set. This is done by consider the frequency of those item A, the frequency of item B, and the co-occur frequency of item A and item B. Association measures are suitable for considering large data set and compare individual items. Association measures express the relationships between item A and item B not just an if-then base but with probabilistic, it uses numbers to express the degree of uncertainty for each item [1, 30, 55].

When we apply the association measures to find relations between items, there are two required steps. The first step is to gather frequencies for item A, item B, or both item A and item B. The second step is to perform association measure calculations to find the relationship between item A and item B by using four types of association measure formulas: Support, Confidence, Lift (interestingness), and Conviction. The details of each one are described below.

The fundamental of the association measures is the Support, it is used for calculating other three association measure formulas. The Support is simply the frequency of an item. The larger the Support scores the higher chance this item exists in the data set.

$$supp(A) = \frac{A}{N} \tag{2.11}$$

where A is the frequency of the item A in the data set, N is the number of total data in the data set.

The first extension from the Support is the Confidence. Confidence represents a chance or certainty that when item A occurs, what is the degree of item B occurs.

$$conf(A \Rightarrow B) = \frac{supp(A \cup B)}{supp(A)} \tag{2.12}$$

where A is the frequency of item A in the data set, $A \cup B$ is frequency when item A and item B co-occur together in the data set.

The Lift or Interestingness [41] can be seen as an extension from Confidence, it takes the correlation between item A and item B. Lift measures the item A and item B in a simple correlation measure. Lift tests on two hypotheses, first the item A does not co-occur with the item B. Second item A and item B are co-occurring together. Lift is the ratio of Confidence to Expected Confidence [29].

$$lift(A \Rightarrow B) = \frac{supp(A \cup B)}{supp(A) \times supp(B)} \tag{2.13}$$

where A is the frequency of item A in the data set, B is the frequency of item B in the data set, $A \cup B$ is the frequency when item A and item B co-occur together in the data set.

Conviction was one of the last defined association measures in 1997, it represents the ratio of the expected frequency that item A occurs without item B. The more independent for the item A and item B the higher the result closes to the maximum 1.

$$conv(A \Rightarrow B) = \frac{1 - supp(B)}{1 - \frac{supp(A \cup B)}{supp(A)}} \tag{2.14}$$

where A is the frequency of item A in the data set, B is the frequency of item B in the data set, $A \cup B$ is the frequency when item A and item B co-occur together in the data set.

2.3.4.5 Machine learning techniques

One of the most important new emerging knowledge in artificial intelligent is machine learning. Machine learning is the process of extracting hidden patterns from the data with machine learning algorithms. Many researches use machine learning as a tool for selecting the correct Chinese OOV translation. Machine learning methods can be classified into 3 levels, they are base learners, meta level learners, and optimizers.

1. Base learners

Base learners are fundaments of machine learning algorithms, they are usually used for classification and regression. Following, we explain some base learner algorithms that are usually used for term translation classification.

a) Decision tree

Decision tree is a supervised classification method. Following we give a brief introduction of decision tree and how it classify items [62].

Decision tree is a well-researched decision support system uses an if-then base and a tree-like graph for grouping each item into different categories. It takes consideration of the each item's features (item's frequency, item's association measure scores, etc) and help to categorize them into positive and negative groups. It is commonly used in decision support system for user to find common features in large data sets. In the tree-like graph, each leaves represent classifications, where each nodes represent joints of features which leads to those class [39, 80].

Decision tree algorithms are very well developed, which are built specifically for classification. We can use the decision tree algorithms to classify different items into correct and wrong categories. Usually it required two steps for decision tree to classify an item. First step is construction of decision tree using training data set. Where the training data set can be built by manually classify few items into correct class and wrong class. Second step is using the constructed tree to predict the correct items from experimental data set [3, 29, 47, 76].

b) Naïve Bayes

Naïve Bayes is a supervised classification method. Following we give a brief introduction of Naïve Bayes and how it classify items

Naïve Bayes assumes each feature are independent, thus the presence or absence of a feature is unrelated to any other features. When training a Naïve Bayes model, it applies the training set to a probability model and obtains class probabilities. When predicting an unknown instance, Naïve Bayes compute the features of such instance into class probabilities. The instance is predicted to the class that has the larger results [23, 49].

c) K-nearest neighbor

K-nearest neighbor is a supervised classification method, following we give a brief introduction of K-nearest neighbor and how it classify instances

K-nearest neighbor predicts the instance according to the distance and votes of its neighbors. K-nearest neighbor is a lazy and simple machine learning algorithm. K-nearest neighbor first train all labeled instances according to their features. Then K-nearest neighbor classify an unknown instance according to the majority of votes of its nearest neighbors (labeled instances). Where k stands for the number of nearest neighbors we take into consider, usually k greater than 1 is preferred for non-contradiction classification [11, 20].

d) Support vector machine

Support vector machine(SVM) is a supervised classification method. Following we give a brief introduction of SVM and how it classify items.

Support vector machine constructs a hyperplane or set of hyperplanes in a high dimensional space. Support vector machine tries to classify the instance by mapping each instance on a hyperplane according to the features of each instance. When training, Support vector machine tries to find the maximum distance between two or more clusters of the labeled instances. Support vector machine try to draw a hyperplane between different classes. When predicting unknown instance, support vector machine map such instance on the hyperplane, support vector machine tries to find which classes does such instance belong according to the hyperplane and the line[19, 57].

e) Artificial neural network

Artificial neural network is a supervised/unsupervised classification method, following we give a brief introduction of artificial neural network and how it classify instances

Artificial neural network is a network of instances that predicts the hidden outcome, the outcome is determined by the connections between the instances and its features. For supervised classification, artificial neural network are given a set of instance pairs and try to find a function which would classify the instance in the correct class. Artificial neural network tries to infer the mapping implied by the instances. Usually a cost function is used to find the errors between the generated mapping and instances. A commonly used cost function is mean-squared error, it tries to minimize the average squared error between the network's output. When an unknown instance is applied to artificial neural network, it will predict the instance according the function [4, 9].

2. Meta level learners

Meta level learners are machine learning methods that use to boost the performance and reduce the problems of base learners. We describe bagging in the following subsection.

a) Bagging

Bagging is a meta-level machine learning method for generating multiple versions of base machine learning models, which it uses, in turn, to get an aggregated predictor model. It separates the input data set into multiple subsets and training multiple base machine learning models. When predicting an unknown instance, bagging validates the instance by the multiple base machine learning models and classifies the instance according to the majority classifications from multiple base machine learning models [10, 64]. For example, the data is separated into 10 folds, of which 9 are assigned for training and one for testing. For each of the training fold, a base machine learning model is generated, which can be Lib-supported vector machine, k-nearest neighbor, etc. After all of the 9 folds have been trained, 9 different base machine learning models are collected, and they can be Lib-supported vector machine, k-nearest neighbor, etc. For testing, each base machine learning model will be applied in order to classify the testing data. The final classification is based on the majority of the classifications obtained from the 9 different base machine learning models. The whole process is conducted 10 times to evaluate all the instances in the data set [10].

3. Optimizations

Machine learning optimization are methods that use to boost the performance and reduce many problems over machine learning errors. We describe some optimizations in the following subsections.

a) Feature selection

When many features are used, it would be important to select the features that can help in classification rather than features than does not help in classification. For a machine learning method, the brute force method is the most straightforward way to find the best combination from all possible combinations of features, but it is known to be time consuming and extremely expensive in the computational cost. There are some feature selection strategies, such as the backward elimination, which can reduce the number of combinations of features to be evaluated and find the best combination of features at the same time. The backward elimination does this by comparing the performances of combinations of features when each different feature is omitted in the full features. Assuming there are 21 features, the backward elimination first generates 21 different sets of features by removing one different feature in each set. Each set has only 20 features now. The set of features that yield the best performance is selected and used as the initial set of features, which will be used by the backward elimination in the new round. The process is repeated while there is still improvement in performance [18].

b) Feature weight optimization

Although some base machine learning algorithms such as Lib-support vector machine can optimize the weight of each feature by themselves, but other base machine learning algorithms such as k-nearest neighbor and Naïve bayes cannot optimize the weight of each feature by themselves. The backward feature weight optimization can optimize the weights of features for base machine learning algorithm. The backward feature weight optimization assumes that features are independent, and it optimizes the weights of each feature with a linear search. Assuming there are 21 features, the backward feature weight optimization generates 11 sets of features with modification of the feature weight (for example, from 0.0 to 1.0) for only feature #21. The set of features that yield the best performance is selected and used as the initial set of features, which will be applied by the backward feature weight optimization in the new round. The process loops for each feature, so that the feature weights with the best performance are selected [36].

c) Parameter optimization

The parameter optimization improves the overall performance of a machine learning algorithm. The evolutionary parameter optimization begins with generating many possible solutions of the parameter settings to form the initial parent solutions. Then the parent solutions are evaluated by a process called fitness ranking. The fitness ranking quantifies the optimal solution (for example, 100% accuracy) and ranks all parent solutions against the optimal solution. Then all parent solutions are grouped in pairs of two according to their ranks (for example, the highest rank and the second highest rank are grouped together). Now the crossover method is applied to each pair to form two new child solutions. The crossover method takes the first half of the solution from parent A and the last half of the solution from parent B to form the first child, then the crossover method takes the last half of the solution from parent A and the first half of the solution from parent B to form the second child. When all the child solutions have been formed, the size of the solutions would be the same as that of the parent solutions. Then a new round of the above process is started with the child solutions. The evolutionary parameter optimization stops when there is no more increase in performance [5].

d) Sampling

One fundamental problem of machine learning is caused by the unbalanced data set. Most data sets in the real world have more incorrect instances than correct ones. We can sample a portion of the incorrect instances according to the size of the correct instances. The sample represents the characteristics of the whole incorrect instances, and is used to be merged with the correct instances to form a new balanced data set. It counts the amount of the correct instances in the data set at first, and then it samples a portion of the incorrect instances with the same amount of the correct ones [27].

2.4 Existing problems for OOV term translation

In section 2.3, we presented many existing OOV term translation methods, however they suffer from several problems. These problems can be listed as follows. First, when the brute force method is used to generate Chinese translation candidates, it may cause the overall translation performance to drop due to too many wrong translation candidates [13, 45, 84-86]; Second, when some popular patterns are used, for example, "Chinese translation (English OOV)" and "[Chinese translation (English OOV)]", to extract a correct Chinese translation, it may reduce the recall, because some correct Chinese translations do not exist in these popular patterns [73]; Third, existing approaches assume that Chinese translation contains only Chinese characters, and ignore any non-Chinese character which may be important in the Chinese biomedical translation [13, 45, 84-86]; Fourth, some existing approaches require human to select the correct translation from a list of translations [85, 86]; Fifth, when supervised machine learning is used for OOV term translation, high cost for feature selection, feature weighting, parameter optimization, and problems for imbalanced data set could result in low performance [71, 89]; and finally, the lexical mapping, word/sub-word-based approaches achieve extremely high precision but gain low recall [81].

Most Chinese translations contain only Chinese characters, however, some Chinese biomedical translations contain non-Chinese characters as a part of the translation, such as Roman alphabets, symbols and Arabic numbers. The existing candidate generation methods ignore those non-Chinese characters, resulting in incorrect results [13, 45, 84-86]. These existing methods generate Chinese translation using only Chinese characters, because it is very difficult to decide which symbols should be included in the Chinese translation. Furthermore, some translations are adjacent to the OOV term, and some translations are disjoint to the OOV term. So the best approach for existing methods is extracting 30 Chinese characters in front and behind the OOV term. Table 2.7 shows the use of only Chinese characters compared with using all kinds of characters. It can be easily seen using only Chinese characters yields a better result. Existing approaches have a large drawback on biomedical type OOV terms especially hybrid translations, for example, “DiGeorge's syndrome” (Human: DiGeorge's 症候群) (existing method: 症候群). Solving hybrid translations is important because existing approaches assumes Chinese translations includes only Chinese characters, which 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 biomedical terms end with the term “症候群”.

Table 2.7: Candidate generation using only Chinese characters compared with using all kinds of characters

English OOVs	**Correct Chinese translations**	**Snippets**	**Using Chinese characters only [1-5]**	**Using all kinds of characters [1-5]**
Cystinosis	胱胺酸病	Urea Cycle ... Cystinosis 胱胺酸病 10.	胱胺酸病	胱胺酸病 10.
GM1/GM2 gangliosidosis	GM1/GM2 神經節苷脂儲積症	GM1/GM2 神經節苷脂儲積症 ICD9-CM. GM1/GM2 gangliosidosis. 疾病名稱	神經節苷脂儲積症	GM1/GM2 神經節苷脂儲積症 ICD9-CM.
Nitroacetylglutamate synthetase deficiency	乙醯穀胺酸合成缺乏症	... Nitroacetylglutamate synthetase deficiency ,NAG synthetase deficiency 乙醯穀胺酸合成缺乏症.	乙醯穀胺酸合成缺乏症	NAG synthetase deficiency 乙醯穀胺酸合成缺乏症
DiGeorge's syndrome	DiGeorge's 症候群	胱胺酸病. 45. DiGeorge's syndrome. DiGeorge's 症候群.	症候群	DiGeorge's 症候群
Correctness	**4/4**		**2/4**	**1/4**

Other fundamental problem of existing approach is the translations mined from the

Internet 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. Retrieval not only translations but also useful knowledge for the end user has become important recently. Most importantly, non-existing method does cross language definition extraction for OOV terms. Moreover, it has always been very difficult for researcher and users to evaluate the correctness of Chinese translations of English OOV terms.

2.5 OOV term definition retrieval

Definition extraction is a subtask of terminology extraction [54]. It is important for natural language processing. It is useful for question answering, automatic creation of glossaries, learning lexical semantic relations and automatic construction of ontologies. Definition extraction for terms aims at finding the definition of such term from books or corpus. Existing methods attempt to extract definitions of terms from domain specific books and corpus [48, 54], they utilize lexicon-syntactic markers and pattern matching to detect definitions. Recent works utilize web data with pattern matching [66]. Most works are targeted at English definition extraction. Little work has been proposed for Chinese language and OOV terms. Zhang and Jiang were pioneers in Chinese term definition extraction and proposed a bootstrapping algorithm for finding term definitions in Chinese language [16]. Their work was targeted at common noun terms rather than OOV terms. However, non-existing methods offer cross language definition extraction.

2.6 OOV terms in CLIR

Cross language information retrieval is an emerging field of study dealing with multilingual translation and information retrieval. It lets the user input queries in a source language to find information written in a target language. It improves information retrieval and knowledge acquiring [60]. CLIR has many useful applications, for example, it assists user to translate words, phrases, and documents from one language to another. Bilingual dictionaries and bilingual translation software are all tools of CLIR, but both above tools cannot perform a high-accuracy bilingual translation without human interferes. Till today, King Soft [37], one of the best bilingual translation software in Chinese to English/English to Chinese, is only able to reach 40% to 50% accuracy when cross language translation is applied with newspaper articles. One of the problems cause such low accuracy is the OOV problem.

Cross language information retrieval aims to retrieve target language documents from source language queries. CLIR employs multilingual dictionaries to achieve such task. OOV term is a large problem for dictionary based CLIR, because multilingual dictionaries have a low inclusion rate of daily emerging OOV terms. Solving OOV term problem would improve the performance of CLIR. Existing method tends to use automatic mining for finding translations of OOV terms and use these translations in the CLIR. Existing CLIR corpus have little inclusion rate of OOV terms, thus it is difficult to evaluate the performance of OOV term translation in CLIR [6, 26, 90].

Chapter 3

Our approaches

In this chapter we describe our approaches for finding translations of both name type and biomedical type OOV terms. Furthermore, we also describe our approaches for finding multilingual context information and monolingual definitions for OOV terms. Moreover we describe our approaches for cross language information retrieval for OOV term definition extraction.

3.1 System design

OOV term translation plays an important role in natural language processing, especially in cross language information retrieval, information retrieval and knowledge acquiring. Out of vocabulary terms are typically new terms that cannot be found in dictionaries. Given an OOV term in source language, OOV term translation extraction aims to find the correct translation in target language. Early solutions for OOV term translation employ multilingual dictionaries. Recent solutions use web-mining together with machine learning. In the past 20 years, many existing methods had explored various ways of finding translations for name type OOV terms in other languages. Yet, none of the existing methods offer detail information and knowledge of OOV terms for the users. Without additional information, the obtained translations are also OOV terms just in other languages. 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. Retrieval not only translations but also useful knowledge for the end user has become important recently. Moreover, existing methods cannot perfectly handle both name type and biomedical type OOV terms, especially hybrid translations, such as “Kenny-Caffey syndrome (Kenny-Caffey 氏症候群)”. Another fundamental problem is many existing methods are able to retrieve English definitions of a term using domain specific books or corpus. However, such methods are not very suitable for OOV terms, because these books or corpus do not contain all of the definitions of daily emerging OOV terms. Furthermore, non-existing methods focus on cross language definition retrieval for OOV terms. Never the less, it has always been so difficult to evaluate the correctness of an OOV term translation without domain specific knowledge and correct references.

We propose a novel translation, information extraction and CLIR for OOV terms. Firstly, this system uses a novel English definition ranking algorithm to differentiate the types of OOV term. Then a novel adaptive rules approach together with supervised machine learning is used for finding translations for biomedical type OOV terms, and a statistical filter together with existing method is used to translate name type OOV terms. Our novel English definition ranking algorithm also extracts multilingual context information and monolingual definitions for name type and biomedical type OOV terms. Moreover, a novel CLIR for Chinese definition ranking algorithm is proposed to extract the Chinese definitions of English OOV terms. Furthermore, we propose a novel auto re-evaluation algorithm to evaluate the correctness of Chinese OOV translations, and Chinese definitions.

For example, "Mae West", our method 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), we would also extract its definition in English (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...), definition in Chinese (梅·蕙絲(英语：Mae West，1893 年 8 月 7 日－1980 年 11 月 22 日)，演員、劇作家、螢幕編劇、也是美國眾所週知的性感偶像。 梅·蕙絲最為人所熟知的是其黃色的雙關語，在到好萊塢為電影產業寫劇本、做諧星、演員之前，她就在紐約的劇場舞台上以綜藝 ...), Chinese definition auto evaluation score of 0.6, and Chinese translation auto evaluation score of 1. A comparison of our method and existing methods is shown in Table 3.1.

Table 3.1: Examples of existing methods and our method

OOV terms	Existing methods	Our method						
	Translation	Translation	Translation correctness	English context information	English definitions	Chinese context information	Chinese definitions	Definition correctness
Mae West	梅蕙絲	梅蕙絲	1	American/ actress/ playwright/ screenwriter/ writer	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...	美国人/ 女演员/ 剧作家/ 编剧家/ 作家	(梅•蕙絲(英语：Mae West，1893 年 8 月 7 日－1980 年 11 月 22 日)，演員、劇作家、螢幕編劇、也是美國眾所週知的性感偶像。 梅•蕙絲最為人所熟知的是其黃色的雙關語，在到好萊塢為電影產業寫劇本、做諧星、演員之前，她就在紐約的劇場舞台上以綜藝 ...)	0.6
α1-Antitrypsin deficiency	抗胰蛋白酶缺乏症	α1-抗胰蛋白酶缺乏症	1	deficiency/ genetic/ disorder/ defective	Alpha 1-antitrypsin deficiency (α1-antitrypsin deficiency, A1AD or simply Alpha-1) is a genetic disorder that causes defective production of alpha 1-antitrypsin ...	缺乏/ 遗传/ 障碍/ 缺陷	α1-抗胰蛋白酶缺乏症是血中抗蛋白酶成份-α1-抗胰蛋白酶（简称α1-AT)缺乏引起的一种先天性代谢病，通过常染色体遗传。临床特点为新生儿肝炎，婴幼儿和成人的肝 ...	0.625
porphyria	紫質症	紫質症	0.2	rare/ inherited/ disorder/ enzyme	The porphyrias are a group of rare inherited or acquired disorders of certain enzymes that normally participate in the production of porphyrins and heme. They ...	稀有/ 遗传/ 障碍/ 酶	紫質症（英語：Porphyria，又稱噗瑳症、卟啉症或吡咯紫質症）是一組因為人體內的紫質（Porphyrin）等物質異常累積所造成的身體 ...	0

The remaining of this chapter is organized as follows: subsection 3.2 introduces the need of OOV type prediction; subsection 3.3 introduces our procedures and flowchart; subsection 3.4 introduces our approach for snippet retrieval; subsection 3.5 introduces English definition extraction; subsection 3.6 introduces multilingual context information extraction and OOV term type prediction; subsection 3.7 introduces translation methods for biomedical type OOV terms; subsection 3.8 introduces translation methods for name type OOV terms; subsection 3.9 introduces CLIR for Chinese definition extraction; and subsection 3.10 introduces translation score extraction.

3.2 The need of OOV type prediction

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 thesis, 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 prediction 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.

3.3 Procedures

The procedures of our approach are as follows:

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 English ranking approach.

3) The multilingual context information extraction and OOV type prediction: ontology tree is constructed from WordNet to extract the context information and predict the type of OOV terms.

4) The biomedical type OOV term translation: biomedical type OOV terms are translated by our novel adaptive rules approach with machine learning.

5) The name type OOV term translation: name type OOV terms are translated by our statistical filter and existing ranking list approach.

6) CLIR for Chinese definition extraction based on generated translations: Chinese definitions are extracted using our novel CLIR for Chinese definition ranking method.

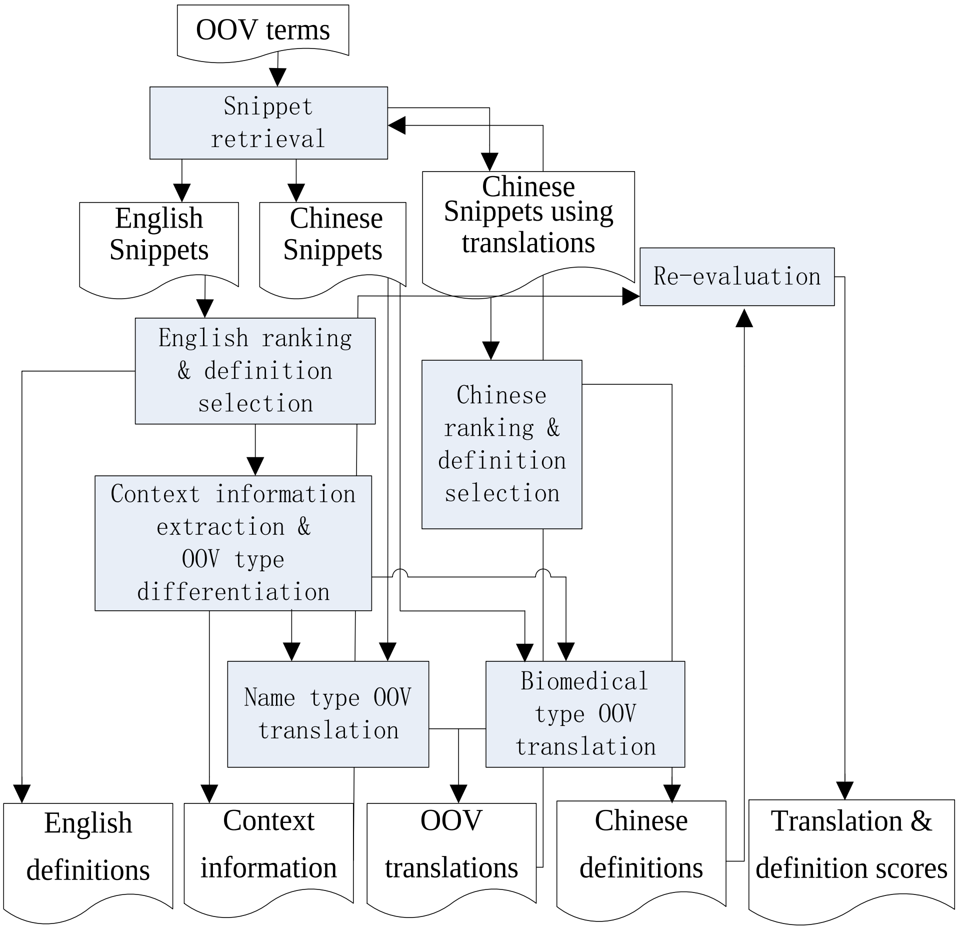

OOV terms
Snippet retrieval
English Snippets
Chinese Snippets
Chinese Snippets using translations
Re-evaluation
English ranking & definition selection
Chinese ranking & definition selection
Context information extraction & OOV type differentiation
Name type OOV translation
Biomedical type OOV translation
English definitions
Context information
OOV translations
Chinese definitions
Translation & definition scores

Figure 3.1: Flow chart of our proposed approach

7) Regression context information evaluation based on Chinese definitions: Chinese translations and definitions are re-evaluated based on Chinese definitions and English context information.

These steps are summarized in the diagram shown in Figure 3.1, the details of each step are described in the following subsections.

3.4 Snippet retrieval

We feed OOV terms to Bing Search API and retrieval snippets from both English and Chinese languages [8]. Furthermore, we also extract the Chinese snippets from Bing

English snippets by English OOV:

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 by English OOV:

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)

Chinese Snippets by Chinese OOV:

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 3.2: An example of web retrieved snippets

Search API by using our selected translations from subsection 3.7 and 3.8. An example of English snippet containing OOV term and its definition; an example of Chinese snippet containing the OOV term and its translation; and an example of Chinese snippet containing our selected translation and its definition are shown in Figure 3.2.

3.5 English definition ranking and definition selection

In order to differentiate the types of OOV terms, we need to firstly extract the English definitions of OOV terms. According to our observation, many OOV terms have their English definitions on the web retrieved English snippets. There are 5 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 used to identify snippets with definitions. 2) Search engines utilize various AI technologies to find related information [59]. Search engine ranks the webpage very carefully, high search engine rank can also led to snippets with definition. 3) Webpages with definitions are usually well organized organization, government, or educational webpages. Thus domain names of the webpages can also led to snippets with definitions. 4) Disambiguation noun terms can simply differentiate definition between an OOV term and a common noun term, for example, Apple (Company) would have terms such as company within the definition, where as Apple (fruit) would have terms such as plant or fruit within the definition. 5) wiki related websites provide high quality definition of some OOV terms, If terms such as wiki, IMDB, and CDC are found in the snippets, it is more like the snippets contain high quality definitions. Combine above five factors, we propose a novel English definition ranking method to rank the snippets with definitions. This method takes snippets retrieved from the Internet and a list of verbs that can identify the snippets with definitions. We consider the location and co-occurrence between the verbs and the OOV terms. Then we combine the domain ranks, search engine ranks, disambiguation noun terms and wiki term ranks to select the snippets with definitions.

The English definition 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 search engine ranks, *DR* be the domain ranks, *N* be disambiguation noun terms and *WR* be the wiki term ranks.

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.

Input: Summaries of snippets retrieved from the Internet, Ss; Titles of snippets retrieved from the Internet, St; OOV terms, OOV; Verb list, V; Search engine ranks, SR; Disambiguation noun list, N; Domain ranks, DR; wiki term ranks, WR.

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,
   If (N found in St) then
   SwR+=1,
   SwD = MaxRank(SwR) && MaxRank( SR) && MaxRank( DR) && MaxRank( WR)
End
```

Figure 3.3: Algorithm English definition ranking

Furthermore, if a disambiguation noun term is found in the *Ss*, we increase the rank with 1.

After the ranks are assigned to the snippets, for each OOV term, we select one snippet with highest English definition rank (max rank), highest search engine rank, highest domain rank, and highest wiki term rank. Domain ranks are given as follows: gov/org/edu/int > com/pro/net/info/ > else. Wiki term ranks are givens as follows: WIKI/CDC > IMDB > else. For each selected snippet, we extract the summary of the snippet as the definition of the OOV term. A detailed algorithm of the English definition ranking is shown in Figure 3.3.

An example of English definition ranking is shown in Table 3.2. As can be seen from this example, the snippet containing the correct definition of the OOV term "Mae West" gained high ranks.

Table 3.2: Ranking results of snippets

URL	Title	Summary	Ranks(SWR, SR, DR,WR)
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 ...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	1,1,1,1
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) · She Done Him Wrong (1933) · My Little Chickadee (1940) · Klondike Annie (1936). Mae West was born in Brooklyn, New York, to ...	1,2,2,2
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?with Cary Grant in this 1933 comedy classic. Fortuneteller: I see a man in your life. Mae: What, only one?	2,3,2,3

3.6 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 WordNet to identify the types of OOV terms. WordNet is a large English lexicon [22]. We use 7 WordNet 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 WordNet super node noun terms (such as illness, sickness, etc.) and extract their brief hyponyms to construct the ontology tree for biomedical type OOV terms. When any word in definition of an OOV term is found in the ontology trees, we identify this OOV term to either name type or biomedical type according to the majority number of words found in different ontology trees. Furthermore, these context information are then translated into Chinese by multilingual dictionary [53] .

3.7 Translation extraction and selection for biomedical type OOV terms

We propose a new adaptive rules system for biomedical type OOV terms in order to handle hybrid translations. The translation extraction and selection for biomedical type OOV term is developed into three steps, they are: firstly, the translation candidate extraction using novel adaptive method; secondly, the feature extraction, and finally the translation selection using machine learning.

3.7.1 Translation extraction for biomedical type OOV terms

The translation candidate extraction is based on the idea that most OOV terms have their correspondent human translation nearby [13, 22, 84]. 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. A flow chart of our translation extraction method is shown in Figure 3.4.

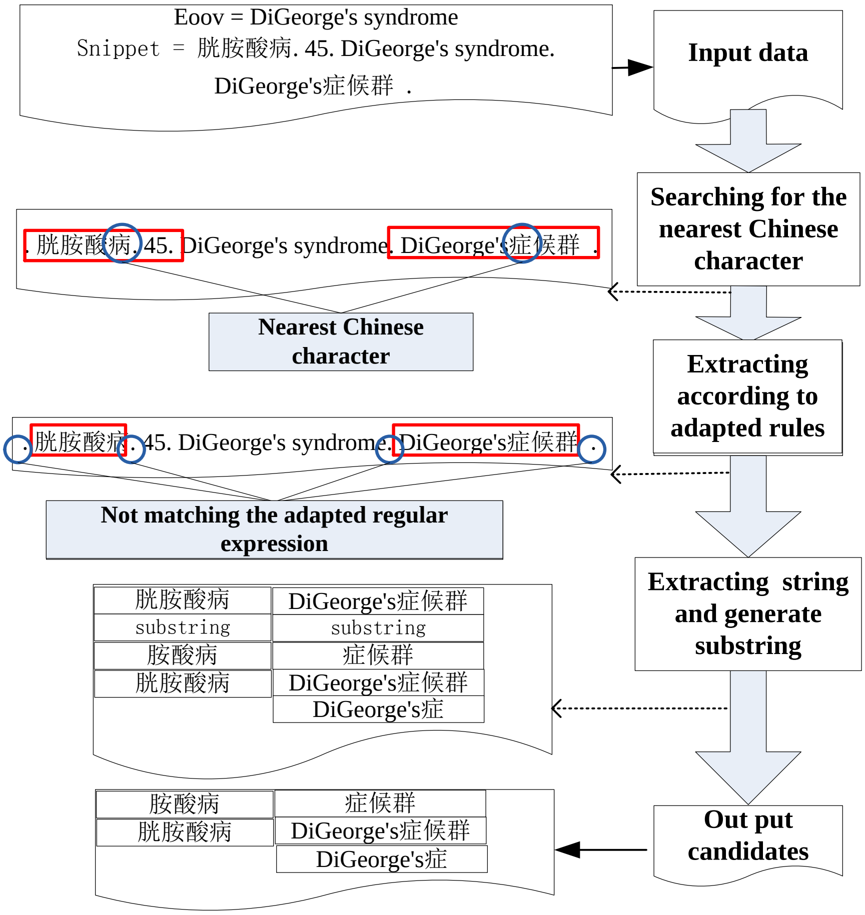

Eoov = DiGeorge's syndrome
Snippet = 胱胺酸病. 45. DiGeorge's syndrome. DiGeorge's症候群 .
Input data
Searching for the nearest Chinese character
. 胱胺酸病. 45. DiGeorge's syndrome. DiGeorge's症候群 .
Nearest Chinese character
Extracting according to adapted rules
. 胱胺酸病. 45. DiGeorge's syndrome. DiGeorge's症候群 .
Not matching the adapted regular expression
Extracting string and generate substring
胱胺酸病
substring
胺酸病
胱胺酸病
DiGeorge's症候群
substring
症候群
DiGeorge's症候群
DiGeorge's症
Out put candidates
胺酸病
胱胺酸病
症候群
DiGeorge's症候群
DiGeorge's症

Figure 3.4: A flow chart of candidate generation for the English OOV term "DiGeorge's syndrome."

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 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 method is explained in Figure 3.5.

```
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
```

Figure 3.5: Algorithm adaptive candidate extraction

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 3.3.

An example of translation candidate extraction is shown in Table 3.4. As can be seen from this example, the OOV term is α1-antitrypsin deficiency, we created the adapted matching rules using the substring of the OOV and the base regular expression matching rules in Table 3.3. The correct translation is extracted as α1-抗胰蛋白酶缺乏症.

Table 3.3: 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 Other characters = all non-Chinese language alphabets, characters, symbols.	

Table 3.4: Example of translation candidate extraction

OOV	α1-antitrypsin deficiency	
Snippet	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。	
Adapted 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	α1-抗胰蛋白酶缺乏症 是以婴儿期出现胆汁	

3.7.2 Feature extraction

In this subsection, we extracted totally 25 different features from OOV terms and translation candidates. They are frequency-related features, distance-related features, a combination of these two groups, and dictionary related feature. The new features proposed in this thesis are marked with novel signs. Details for each group are explained below.

3.7.2.1 Frequency-related features

The frequency represents a very important statistical feature for the retrieved Chinese translation candidates. The more frequently a Chinese translation co-occurs with an English OOV term, the more likely it is the correct translation. The frequency-related features consist of 12 major features that are related to frequency, they are individual frequencies of the Chinese translation and the English OOV term, the co-occurrence frequencies of the Chinese translation and the English OOV term, the number of snippets of each English OOV term, the Symmetrical conditional probability, chi-square, support, lift, confidence and conviction. Details of each feature are explained below.

1. **[f1-f5] Frequencies ($\boldsymbol{f(c_i)}$, $\boldsymbol{f(e_i)}$, $\boldsymbol{f(e_i, c_i)}$, $\boldsymbol{Ff(c_i e_i)}$, $\boldsymbol{Bf(e_i c_i)}$)**

We collect the frequency of the Chinese translation candidates ($f(c_i)$), English OOV term ($f(e_i)$) and their co-occurrence frequency ($f(e_i, c_i)$). In addition, as the Chinese translation candidate occurs either in front or in the back of the English OOV term, we further determine how frequently the translation appears at the front or the back by collecting the front co-occurrence frequency ($Ff(c_i e_i)$) and back co-occurrence frequency ($Bf(e_i c_i)$). These frequencies ($f(c_i)$), ($f(e_i)$), ($f(e_i, c_i)$), ($Ff(c_i e_i)$), and ($Bf(e_i c_i)$) are our first five features.

2. [f6] Symmetrical Conditional Probability ($\boldsymbol{SCP(c_1 \dots c_n)}$)

In some scenario, the Chinese translation obtained from the previous step contains the correct translation with the translation of the neighboring word. To ensure that the candidates are the translation of the OOV by itself, we calculate the Symmetrical Conditional Probability (SCP) [45] score for each translation candidate. The $SCP(c_1 \dots c_n)$ checks each alphabet, character and substring in the possible translation. By calculating the frequencies 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 [50].

$$\boldsymbol{SCP}(c_1 \dots c_n) = \frac{(n-1)f(c_1 \dots c_n)^2}{\sum_{i=1}^{i=n-1} f(c_1 \dots c_i) f(c_{i+1} \dots c_n)} \tag{3.1}$$

where $(c_1 \dots c_n)$ is any possible Chinese translation candidate, n is the number of characters in this Chinese translation candidate, $f(c_1 \dots c_n)$is the frequency of that candidate occurring in the retrieved snippets, and $f(c_1 \dots c_i)$ and $f(c_{i+1} \dots c_n)$ are the frequencies of substrings of the Chinese translation candidate occurred in the retrieved snippets. SCP is our sixth feature used in the experiment.

3. [f7] Chi-square ($\boldsymbol{\chi^2(e_i, c_i)}$)

Our seventh feature is the *Chi-square*. The *Chi-square* [12, 13] tests a list of possible translation candidates with their source OOV term to find a correlation relationship. A numerical correlation relationship between the English OOV term e_i and its Chinese translation candidates c_i can be measured with this method.

$$\chi^2(e_i, c_i) = \frac{N \times (S(e_i \wedge c_i) \times S(\neg e_i \neg c_i) - S(e_i \wedge \neg c_i) \times S(c_i \wedge \neg e_i))^2}{(S(e_i \wedge c_i) + S(e_i \wedge \neg c_i)) \times (S(e_i \wedge c_i) + S(c_i \wedge \neg e_i)) \times (S(e_i \wedge \neg c_i) + S(\neg e_i \neg c_i)) \times \left(S(c_i \wedge \neg e_i) + S(\neg e_i \neg c_i)\right)} \tag{3.2}$$

where e_i is the English OOV term, c_i is the Chinese translation, N is the total number of snippets in the corpus, and S is a function that takes a query as the input and returns the number of snippets from the corpus. The higher the chi-square score, the higher the co-occurrence relationship between the English OOV term and the Chinese translation candidate.

4. [f8-f11] Modified Association Measures ($Supp(e_i \rightarrow c_i), conf(e_i \rightarrow c_i), lift(e_i \rightarrow c_i), conv(e_i \rightarrow c_i)$) (novel)

We propose the modified association measures, which do not require the total number of pages in the Internet. They take the webpage counts of OOV terms $S(e_i)$, translations $S(c_i)$, the webpage counts of OOV terms co-occur with translations $S(e_i \wedge c_i)$ and the webpage counts 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.

$$Supp(e_i \rightarrow c_i) = S(e_i \wedge c_i) \tag{3.3}$$

$$Conf(e_i \rightarrow c_i) = \frac{S(e_i \wedge c_i)}{S(e_i)} \tag{3.4}$$

$$lift(e_i \rightarrow c_i) = \frac{S(e_i \wedge c_i)}{S(e_i)S(c_i)} \tag{3.5}$$

$$Conv(e_i \rightarrow c_i) = \frac{S(e_i)(\neg c_i)}{S(e_i \wedge \neg c_i)} \tag{3.6}$$

where e_i is the English OOV term, c_i is the Chinese translation, N is the total number of snippets in the corpus, and S is a function that takes a query as the input and returns the number of snippets from the Internet.

5. [f12] Number of Snippets ($\boldsymbol{Sn(e_i)}$) (novel)

Each English OOV term retrieves a different number of snippets from the Internet. The more are snippets containing an English OOV term, the higher is the frequencies of its Chinese translation candidates. We collect the number of snippets for each English OOV term. This feature distinguishes the English OOV terms with high frequencies from the ones with very low frequencies. Thus, machine learning will not classify all of the Chinese translations with low frequencies as the incorrect translation.

3.7.2.2 Distance-related features

According to the work of Cheng *et al.* [13] and Zhang *et al.* [84], when a Chinese translation candidate appears close to its input English OOV term, it is more likely that the Chinese translation is correct [13, 84]. We thus consider the distance as one of our features. The distance-related features contains seven features, they are the front average, back average and total average distance between the Chinese translation and the English OOV term, candidate length, OOV length, difference between them, and length similarity. Details of each feature are explained below.

1. [f13-f15] Front average, back average and total average distances ($\boldsymbol{F\overline{d}}(\boldsymbol{c_i})$, $\boldsymbol{B\overline{d}}(\boldsymbol{c_i})$, $\boldsymbol{FB\overline{d}}(\boldsymbol{c_i})$) (novel)

The distance between the locations of the English OOV term and each Chinese translation candidate can be calculated by counting the number of characters separating the Chinese translation candidate from the English OOV term in the web retrieved snippets. 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 front average distance $\boldsymbol{F\overline{d}}(c_i)$, back average distance ($\boldsymbol{B\overline{d}}(c_i)$), and the total front and back average distance ($\boldsymbol{FB\overline{d}}(c_i)$). Formulas for calculating these distances are shown below.

$$\boldsymbol{F\overline{d}}(c_i) \ = \frac{\sum_{k=1}^{k=\mathbf{F}\boldsymbol{f}\boldsymbol{c_i}\boldsymbol{e_i}} |Fd_k|}{\boldsymbol{Ff}(\boldsymbol{c_i}\boldsymbol{e_i})} \tag{3.7}$$

$$\boldsymbol{B\overline{d}}(c_i) \ = \frac{\sum_{k=1}^{k=\mathbf{B}\boldsymbol{f}\boldsymbol{c_i}\boldsymbol{e_i}} |Bd_k|}{\boldsymbol{Bf}(\boldsymbol{e_i}\boldsymbol{c_i})} \tag{3.8}$$

$$\boldsymbol{FB\overline{d}}(c_i) \ = \frac{(\sum_{k=1}^{k=\mathbf{F}\boldsymbol{f}\boldsymbol{c_i}\boldsymbol{e_i}}|Fd_k|) + (\sum_{k=1}^{k=\mathbf{B}\boldsymbol{f}\boldsymbol{c_i}\boldsymbol{e_i}} |Bd_k|)}{\boldsymbol{Ff}(\boldsymbol{c_i}\boldsymbol{e_i}) + \boldsymbol{Bf}(\boldsymbol{e_i}\boldsymbol{c_i})} \tag{3.9}$$

where c_i is the Chinese translation, $Fd_k$is the distance between the English OOV term and the Chinese translation which occurs in front, Bd_k is the distance between the English OOV term and the Chinese translation which occurs in the back. $\boldsymbol{Ff}(\boldsymbol{c_i}\boldsymbol{e_i})$ is the frequency when the Chinese translation occurs in the front of the English OOV term, and $\boldsymbol{Bf}(\boldsymbol{e_i}\boldsymbol{c_i})$ is the frequency when the Chinese translation occurs in the back of the English OOV term.

2. **[f16-f18] Length of OOV and translation candidates ($|\boldsymbol{c_i}|, |\boldsymbol{e_i}|, \mathbf{D}(|\boldsymbol{e_i}|, |\boldsymbol{c_i}|)$)**

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 $\mathrm{D}(|e_i|, |c_i|)$. The candidate length ($|C_i|$) is simply the number of characters in the translation candidates. The correct biomedical OOV term translations are usually long candidates [5], because Chinese translations of biomedical words are usually formed by four or more Chinese characters.

3. **[f19] Length similarity ($\boldsymbol{\delta}(|\boldsymbol{e_i}|, |\boldsymbol{c_i}|)$)**

Other than simple differences between the lengths of OOV and the translations, we also employ the length similarity ratio, it is a normalized length difference $\boldsymbol{\delta}(|e_i|, |c_i|)$. and it is computed as follow [65].

$$\boldsymbol{\delta}(|e_i|, |c_i|) = \frac{|c_i| - |e_i| \times c}{\sqrt{(|e_i| + 1)\sigma^2}} \tag{3.10}$$

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

3.7.2.3 Combination of frequency and distance-related features

The correlation relationship between the frequency, length and distance of the translation candidate and its English OOV term helps us to identify which translation is correct [86]. The combination of frequency and distance-related features contains 5 features, which are the ranking function, fuzzy ranking function, ranking function over distance, co-occurrence distance and modified co-occurrence distance. Details of each feature are explained below.

1. [f20] Ranking function ($\boldsymbol{R}(\boldsymbol{c_i})$)

As the candidate that appears frequently (large$f(c_i)$) and is long (large$|C_i|$) is getting a higher rank than the candidates that rarely appear and are short, this formula is built specifically for OOV term translations, these translations are usually long and appear frequently [86]. We use this idea to calculate the ranking function for each candidate. The formula for calculating the ranking function is shown below.

$$\boldsymbol{R}(c_i) \ = \alpha \ \frac{|c_i|}{\max_j |c_j|} + (1-\alpha)\frac{f(c_i)}{f(e_i)} \tag{3.11}$$

where $\max_j |c_j|$ is the maximum length of the Chinese translation candidate found, $|c_i|$ is the length of each possible Chinese translation candidate, $f(c_i)$ is the frequency of the Chinese translation candidate, and $f(e_i)$ is the frequency of the English OOV term. α is a weighting constant. Zhang and Vines [86] suggested that α = 0.25 is recommended as it provides the best combination of frequency and length. α is used for balancing the importance of the length and frequency of the candidates.

2. [f21] Fuzzy Ranking function ($\boldsymbol{FR}(\boldsymbol{c_i})$) (novel)

From the previous method, finding the appropriate α is not an easy task, therefore we introduce a fuzzy ranking function which employs $SCP(c_1 \dots c_n)$ to be used instead of α. The formula for calculating the fuzzy ranking function is shown below.

$$\boldsymbol{FR}(c_i) \ = SCP(c_1 \dots c_n) \times \frac{|c_i|}{\max_j |c_j|} + (1 - SCP(c_1 \dots c_n)) \times \frac{f(c_i)}{f(e_i)} \tag{3.12}$$

where $SCP(c_1 \dots c_n)$ is the Symmetrical Conditional Probability of the Chinese translation candidate c_i.

3. [f22] Ranking over distance ($\boldsymbol{RD}(\boldsymbol{c_i})$) (novel)

Because the ranking function finds a correlation between the length and frequency of a Chinese translation candidate, we modify it so that it also considers the average distance between the English OOV term and the Chinese translation. The smaller the distance between the English OOV term and the Chinese translation, the higher the probability the translation is correct. The ranking over distance gives a higher rank to candidates that appear frequently, are long and have low value of the average distance between the OOV term and the Chinese translation candidates.

$$\boldsymbol{RD}(c_i) = \frac{\boldsymbol{R}(c_i)}{\boldsymbol{FB\bar{d}}(c_i)} \tag{3.13}$$

where $\boldsymbol{FB\bar{d}}(c_i)$ is the total average distance between the English OOV term and Chinese translation candidates. $\boldsymbol{R}(c_i)$ is the ranking function of c_i.

4. [f23-24] Co-occurrence distance ($\boldsymbol{CDist}, \boldsymbol{CwDist}$)

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.

$$CDist = \frac{sum(Dist(c_i e_i))}{tf(c_i e_i)} \tag{3.14}$$

A modification of this feature ($CwDist$) was proposed by Zhang *et.al* [88], they use the web retrieved page count instead of the $tf(c_i e_i)$. It is computed as follows.

$$CwDist = \frac{sum(Dist(c_i e_i))}{S(c_i)} \tag{3.15}$$

where $tf(c_i e_i)$ is the co-occur frequency between OOV and translation candidate, $S(c_i)$ is the web retrieved page count of translation candidate.

3.7.2.4 Dictionary related feature

Some researcher argue dictionaries can be used to solve OOV term translation problems [81]. We proposed a dictionary feature as follows.

1. **[f25] Dictionary candidate score *(DIC_i)* (novel)**

Lexical mapping and word/sub-word OOV term translation method was proposed by Zhang and Sumita [81]. We use an English-Chinese dictionary from LDC [43], The dictionary contains 223,862 unique English words with 128,366 unique Chinese translations, some English words having possibly the same Chinese translations. For each OOV term, we split it into English sub-word and try to find the Chinese translation in the dictionary. If any sub-word in an OOV term were found in the dictionary, we combine the Chinese translations from the dictionary into a string and call it the dictionary translation (Dt_i). Finally, we acquire the string similarity score between dictionary translations and our translation candidates.

$$\boldsymbol{DIC}_i = \frac{\#ch(Dt_i \wedge c_i)}{\#ch(Dt_i)} \tag{3.16}$$

where $\#ch(Dt_i \wedge c_i)$ is the number of same characters in both translation candidate and dictionary translation. $\#ch(Dt_i)$ is the number of characters in dictionary translation.

3.7.3 Translation candidate selection for biomedical type OOV terms

In this section, we explain our candidate selection method. It is developed into two parts, the statistical filter, and machine learning.

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

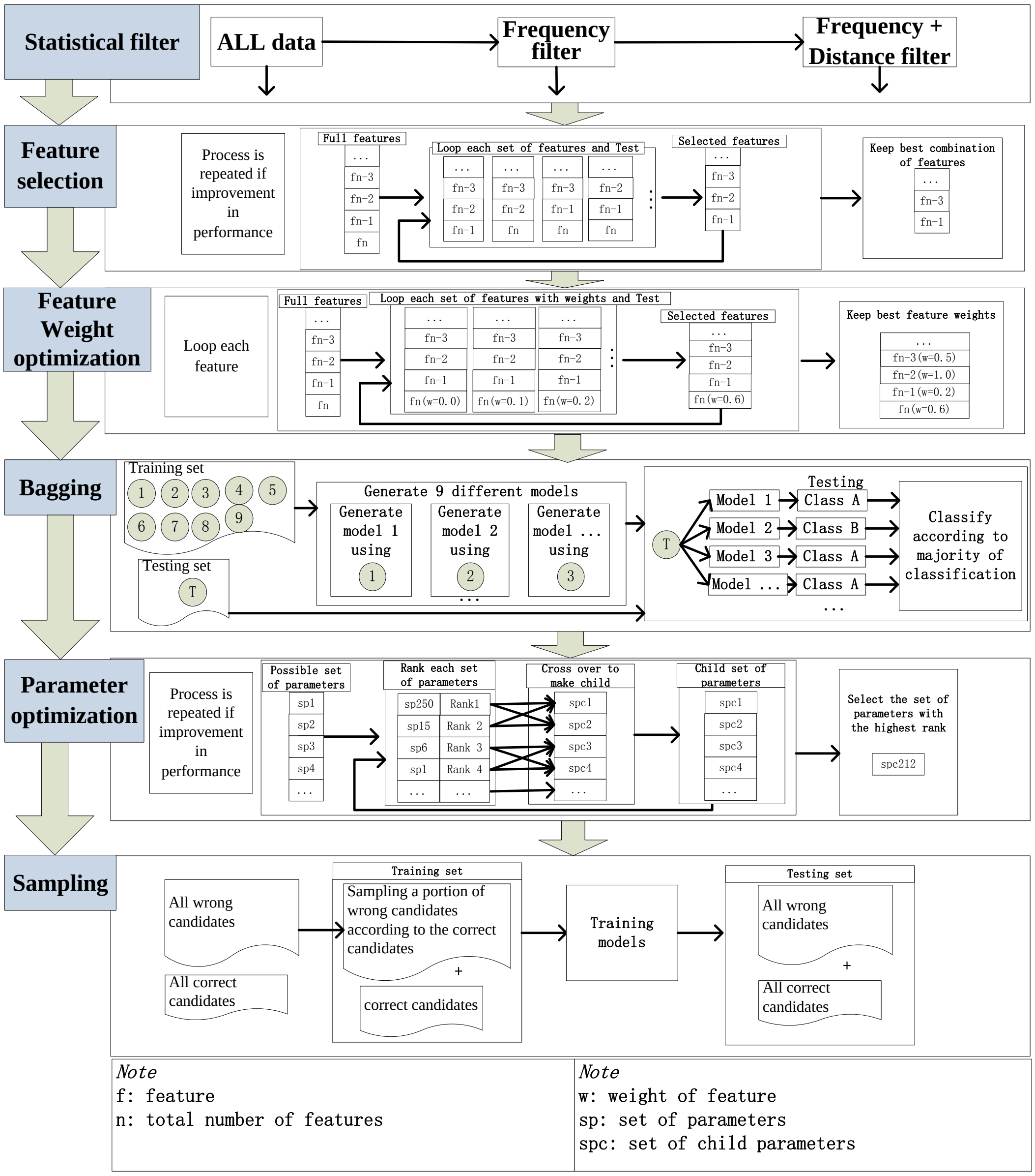


Figure 3.6: Process chart of the Machine learning setup

the correct translation. Two features in our feature set can simply filter some wrong candidates, they are co-occur frequency ($tf(c_ie_i)$) and location distance ($Dist(c_ie_i)$). 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. Our statistical filter takes the top 70% of the co-occur frequency and the shortest location distance between OOV and the translations.

We applied four base level machine learning algorithms which are ANN, LibSVM, k-NN and Naïve Bayes, with novel statistical filter, feature selection, feature weight optimization, bagging, parameter optimization, and sampling.

The optimization processes (such as feature selection, parameter optimization, and feature weight optimization) are only applied to the filtered data set (data after frequency and distance filter) with 10-fold cross validation. Then the settings obtained from the optimization processes are used for larger data sets (data without filter, and data with frequency filter). This setup easily evaluates the ability of our proposed statistical filter. If the obtained optimization settings were useful, they would suggest that our filter is quite able to select a small portion of the data to represent the whole data set. A Figure of our machine learning process chart is shown in Figure 3.6.

3.8 Translation extraction and selection for name type OOV terms

The name type OOV terms are translated by our statistical filter presented in subsection 3.7.3 together with existing ranking list method. The existing ranking list method is based on the idea that most OOV terms co-occur with their correspondent human translations in the web retrieved snippets [13, 22, 84].

For each name type OOV term, we extract 30 Chinese characters before and after the OOV term when source OOV term is found in the snippet. Then, all possible substrings are generated from the extracted Chinese characters. We perform our statistical filter to reduce the number of substrings. Finally, a ranking list method from Zhang and Vines is used to rank the substrings (possible translation candidates). As the candidate that appears frequently (large$f(c_i)$) and is long (large$|c_i|$) is getting a higher rank than the candidates that rarely appear and are short, this formula is built specifically for name type OOV term translations, these translations are usually long and appear frequently with the

English OOV term [86]. The formula for calculating the ranking list method is shown below. Only the top ranked candidate is selected as the Chinese translation.

$$\boldsymbol{R}(c_i) \ = \alpha \ \frac{|c_i|}{\max_j |c_j|} + (1-\alpha)\frac{f(c_i)}{f(e_i)} \tag{3.17}$$

where $\max_j |c_j|$ is the maximum length of the Chinese translation candidate found, $|c_i|$ is the length of each possible Chinese translation candidate, $f(c_i)$ is the frequency of the Chinese translation candidate, and $f(e_i)$ is the frequency of the English OOV term. α is a weighting constant. Zhang and Vines [86] suggested that α = 0.25 is recommended as it provides the best combination of frequency and length. α is used for balancing the importance of the length and frequency of the candidates.

3.9 CLIR for Chinese definition extraction based on generated translations

In this step, we use the obtained Chinese OOV translations from subsection 3.7 and 3.8, and apply to our novel CLIR for Chinese definition ranking algorithm. Chinese language unlike English, it has many special characteristics, the most important difference in natural language processing is word segmentation. Chinese language does not have word boundaries. Another important difference is that many Chinese language characters share same sounds. Moreover, Chinese OOV translations are not unified, especially for terms that using transliteration method. Thus there may be few correct Chinese translations for an English OOV term available on the Internet. Each Chinese translation may use characters in similar sounds, but they are different in writing, for example, a name type OOV term "Benz" with its un-unified Chinese translations"奔驰" and "賓士". Furthermore, Chinese OOV translation does not suffer much in disambiguation like English. “Apple” would be translated into 苹果电脑 not 苹果. Never the less, simply adapt English definition ranking algorithm result in low recall.

In order to solve the above problems, we propose a novel CLIR for Chinese definition ranking algorithm. It utilize Chinese verbs, location and co-occurrence between the verbs and the OOV translations, together with Chinese domain ranks, search engine ranks, Chinese wiki terms and segmentation algorithm to select the Chinese snippets with definitions. According to our observation, Chinese language unlike English language, it uses a larger number of verbs for indicting meaning and definition, such as 為/簡稱/俗稱 are not commonly used in English for indicting definitions. We also employed a n-gram with dictionary based algorithm for Chinese word segmentation to obtain the substrings of the Chinese translations [77].

The CLIR for Chinese definition ranking is developed as follows. Let *Ss* be the summaries of Chinese snippets retrieved from the Internet, *St* be the titles of Chinese snippets retrieved from the Internet, *OOVT* be the Chinese translation of OOV terms, *V* be the Chinese verb list, *SR* be the search engine ranks, *DR* be the domain ranks, and *WR* be the wiki ranks,

For each selected OOV translation:

if the OOV translation and a verb in *V* are both found in the *Ss*, we give it a rank 1;

if the OOV translation is found in the *St* and a verb in *V* are found in the *Ss*, we give it a rank 2;

if the sub-word of the OOV translation and a verb in *V* are found in the *Ss*, we give it a rank 3;

if the OOV translation is found in the *Ss*, we give it a rank 4;

if only sub-word of the OOV translation is found in the *Ss*, we give it a rank 5;

if no sub-word of the OOV term is found in the *Ss*, we give it a rank 6.

After the ranks are assigned to the snippets, for each OOV term, we select one snippet with highest Chinese definition rank (max rank), highest search engine rank, highest domain rank, and highest wiki term rank. Chinese domain ranks are given as follows: gov/org/edu/int/com > /pro/net/info/ > else, because the largest Chinese encyclopedia

```
Input: Summaries of snippets retrieved from the Internet, Ss; Titles of snippets retrieved
from the Internet, St; OOV translations, OOVT; Verb list, V; Search engine ranks, N; Do-
main ranks, DR; wiki ranks, WR.

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 and V found in Ss)then
         SwR = 2,
                Else If (sub-word OOV and V found in Ss)then
                SwR = 3,
                       Else If (OOV found in Ss )then
                       SwR = 4,
                               Else If (sub-word OOV found
                               in Ss )then
                               SwR = 5,
                                     Else
                                     SwR =6,
  SwD = MaxRank(SwR) && MaxRank( SR) && MaxRank( DR) && MaxRank( WR)
End
```

Figure 3.7: Algorithm CLIR for Chinese definition ranking

Baidu uses .com as domain. Chinese wiki term ranks are givens as follows: 百科>知識/知道> else, these is no IMDB Chinese site. For each selected snippet, we extract the summary of the snippet as the definition of the OOV term. A detailed algorithm of the CLIR for Chinese definition ranking is shown in Figure 3.7.

An example of CLIR for Chinese definition ranking is shown in Table 3.5. As can be seen from this example, the snippet containing the correct Chinese definition of the OOV translation "沃尔玛" gained high ranks.

Table 3.5: Ranking results of snippets

URL	Title	Summary	Ranks(SWR, SR, DR,WR)
http://baike.baidu.com/view/9389.htm	沃尔玛百货有限公司_百度百科	沃尔玛公司（Wal-Mart Stores, Inc.）（NYSE：WMT）是一家美国的世界性连锁企业，以营业额计算为全球最大的公司，其控股人为沃尔顿家族。总部位于美国阿肯色州的本顿维尔。	1,1,1,1
http://zh.wikipedia.org/zh-cn/%E6%B2%83%E5%B0%94%E7%8E%9B	沃尔玛 - 维基百科，自由的百科全书	沃尔玛公司（ 英语 ：Wal-Mart Stores, Inc. ）（ NYSE ： WMT ）是一家美国的跨国零售企业，总部设在阿肯色州本顿维尔。为全球最大的公司（以营业额计算）） 。也是世界上最大的私人雇主，员工 ...	1,2,1,1
http://zh.wikipedia.org/zh-cn/Talk:%E6%B2%83%E5%B0%94%E7%8E%9B	讨论:沃尔玛 - 维基百科，自由的百科全书	这是一个讨论关于 沃尔玛 条目相关更改的 讨论页 。 请勿 将讨论页当成讨论这个主题的 论坛 。 请在您发表的意见末加上四条半角波浪号（ ~~~~ ） 以添上 附有时间的签名 。 新的发言置于旧 ...	2,3,1,1

3.10 Regression translation auto re-evaluation

It has always been very difficult for researcher and users to evaluate the correctness of Chinese OOV term translation. In this finally step, we propose a novel regression algorithm to automatically evaluate the correctness of the Chinese translation and Chinese definitions. This algorithm starts with two hypnosis, 1) We assume most English definitions are correct, thus their context information are also correct. 2) We assume a correct

Chinese translation will extract correct Chinese definition and vice versa.

Since we obtained the Chinese context information by machine translation, we can search these context information against Chinese definition. If they can be found within the Chinese definition, it is most likely such Chinese definition is correct, thus Chinese translation is also correct. Detailed algorithm of regression auto re-evaluation as follows:

For each selected Chinese definition, we use its Chinese context information to search against the Chinese definition:

If a full string of a context information is found in definition, we give a score of 1;

If a substring of a context information is found in definition, we give a score of 0.5;

If a context information is not found in definition, we give a score of 0;

Finally, we normalize the score of auto re-evaluation with the number of context information.

Moreover, if a translation generates a definition with a definition evaluation score of more than 0.5, we give 1 for the translation score; if definition evaluation score is less than 0.5, we give a 0.5 for translation score; if definition evaluation score is 0, we give a 0.2 for translation score.

Chapter 4

Experiments and results

In this chapter, we describe the data collection, experimental setup and experimental results for all the approaches we proposed in chapter 3. The rest of the chapter is organized as follows: in subsection 4.1, we describe the evaluation scheme; in subsection 4.2, we describe the data collection; in subsection 4.3, we describe the experimental setup and results for English definition extraction and multilingual context information extraction; in subsection 4.4 we describe the experimental setup and results for OOV term type prediction; in subsection 4.5, we describe the experimental setup and results for name type OOV term translation; in subsection 4.6, we describe the experimental setup and results for biomedical OOV term translation; in subsection 4.7 we describe the experimental setup and results for CLIR for Chinese definition extraction; and finally in subsection 4.8 we describe the experimental setup and results for auto re-evaluation.

4.1 Evaluation scheme

For translation, we manually queried the Internet with English OOV terms and retrieved the correct Chinese translations. We cross-checked those correct translations with few published lists of OOV translations to ensure their correctness [34].

For multilingual context information, and multilingual definitions, we hired two graduate students to check our extraction results.

In order to compare our approach with existing approaches, we tested the same data set with the local maximum and SVM method from Zhang *et al*. [88]. We also tested the length & co-occurrences method from Zhang and Vines [86], and the word/sub-word method proposed by Zhang and Sumita [81]. We use precision, recall, F-measure, accuracy, and blue score to compare our results with existing results. Calculations for precision, recall, F-measure, accuracy, and blue score are shown below:

$$Accuracy = \frac{correct\ translations\ mined\ in\ OOV\ terms}{total\ correct\ translations\ in\ OOV\ terms} \tag{4.1}$$

$$Fmeasure = \frac{2\ \times Precision\ \times Recall}{Precision + Recall} \tag{4.2}$$

$$BLEU = \frac{extracted\ translation\ in\ characters\ \cap\ correct\ translation\ in\ characters}{correct\ translation\ in\ characters} \tag{4.3}$$

$$Recall = \frac{relevant\ documents\ \cap\ retrieved\ documents}{relevant\ documents} \tag{4.4}$$

$$Precision = \frac{relevant\ documents\ \cap\ retrieved\ documents}{retrieved\ documents} \tag{4.5}$$

$$Precision\ for\ classification = \frac{True\ positive}{True\ positive + False\ positive} \tag{4.6}$$

$$Recall\ for\ classification = \frac{True\ positive}{True\ positive + False\ negetive} \tag{4.7}$$

$$Accuracy\ for\ classification = \frac{True\ positive + True\ negetive}{True\ positive + False\ positive + True\ negetive + False\ negetive} \tag{4.8}$$

4.2 Data collection

We collected English name type OOV terms from lists of top 500 companies [17], famous people [63], car brands [75], Furthermore, we collected English biomedical terms from Classification of Diseases, Functioning, and Disability ICD9 [33]. Combining the above lists, we obtained a total of 3,487 OOV terms, 746 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 method on both name type and biomedical type OOV terms.

According to subsection 3.4, we need to collect both English and Chinese snippets for different purposes, the English snippets are used for definition extraction, context information extraction and OOV type prediction; while Chinese snippets are used for translation extraction. We retrieved a total of 319,459 English snippets, and 275,639 Chinese snippets after querying each English OOV terms to the web. Then we processed the English snippets with our English ranking method and the Chinese snippets with our translation extraction method.

4.3 Experimental setup and results for multilingual context information extraction and English definition extraction

Although the input were 3,487 OOV terms, only 2,498 were able to be retrieved form the Internet, 746 of them are name type OOV terms, and 1,752 are biomedical type OOV terms. Our English ranking method extracted the English definition and multilingual context information of the English OOV term. We hired human to check the extracted definition and context information for correctness. Our method is able to retrieve correct English definitions for 1,885 OOV terms, 659 of them are name type OOV terms, and 1,226 of them are biomedical type OOV terms. Moreover, 2,102 of the OOV terms were able to obtain the correct multilingual context information, 718 of them are name type OOV terms, and 1,384 of them are biomedical type OOV terms. The detailed results are

Table 4.1: Experimental results for English definition extraction

	English definition (accuracy)
All OOV(2,498)	1,885(75.46%)
Name type OOV(746)	659(88.33%)
Biomedical type OOV(1,752)	1,226(69.97%)

Table 4.2: Experimental results for English definition extraction without web features

	English definition (accuracy)
All OOV(2,498)	1,565(62.65%)
Name type OOV(746)	536(71.85%)
Biomedical type OOV(1,752)	1,029(58.73%)

Table 4.3: Experimental results for multilingual context information extraction

	English context information (accuracy)	Chinese context Information (accuracy)
All OOV(2,498)	2,102(84.15%)	2,102(84.15%)
Name type OOV(746)	718(96.26%)	718(96.24%)
Biomedical type OOV(1,752)	1,384(79.00%)	1,384(79.00%)

shown in Table 4.1 and 4.3. Furthermore, we tested English definition extraction without using web features, such as domain ranking, searching engine ranking, and wiki terms ranking. The details are shown in Table 4.2.

4.4 Experimental setup and results for OOV type prediction

Our novel English ranking method predicted 934 OOV terms as name type OOV terms, and 1,563 OOV terms as biomedical type OOV terms. The detailed results are shown in Table 4.4.

Table 4.4: Experimental results for OOV type prediction

	precision	recall
All OOV(2,498)		
Name type OOV(746)	79.76%	99.86%
Biomedical type OOV(1,752)	99.93%	89.21%

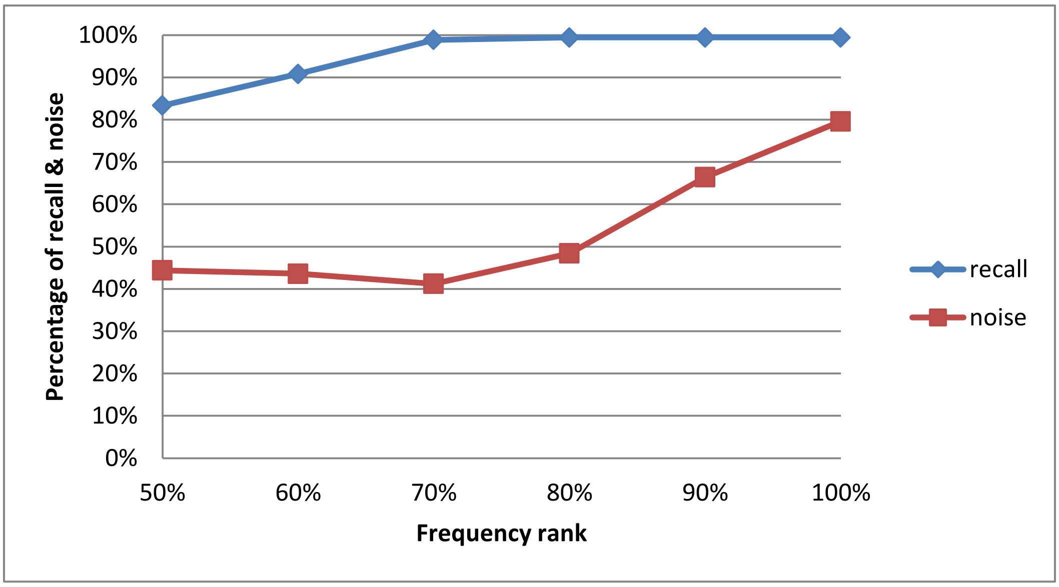


Figure 4.1: Recalls and noises using different frequency rank

4.5 Experimental setup and results for name type OOV term translation

After we predicted the types of OOV terms, the name type OOV terms are translated using our statistical filter with ranking list method and biomedical type OOV terms are translated using our novel adaptive rules and machine learning.

For name type OOV terms, existing ranking list candidate generation method is applied, however we add our novel statistical filter to increase the performance.

We use our novel statistical filter to filter possible translation candidates. We take the top 70% of the frequency rank and select the Chinese translation candidates, which are

Table 4.5: Comparison of our proposed method with existing methods

	Our method	local maximum SVM method	Length & co-occurrences
Transliteratable Name type OOVs	746	746	746
Correct translation mined	736	734	736
Correct translation mined in OOV terms (accuracy)	734 (98.39%)	729 (97.72%)	734 (98.39%)
Correct translation selected in OOV terms (accuracy)	734 (98.39%)	709 (95.04%)	701 (93.96%)

closest in location to their corresponding English OOV term, because top 70% present the best combination of high recall and low noise where noise shows the number of wrong translations included in the candidate set. Figure 4.1 shows recalls and noises using different frequency rank for name type OOV terms.

Table 4.5 shows the experimental results for name type OOV term translation. We can see that length & co-occurrences method from Zhang and Vines gained accuracies of 98.39% and 93.96% in candidate generation and candidate selection respectively. While local maximum and SVM method from Zhang *et al.* gained accuracies of 97.72% and 95.04% in candidate generation and candidate selection respectively. Our proposed methods are slightly better than existing methods. We achieved accuracies of 98.39% and 98.39% in candidate generation and candidate selection respectively.

4.6 Experimental setup and results for biomedical type OOV term translation

For biomedical OOV term, we processed each snippet by our candidate generation method, and we generated 157,075 Chinese translation candidates. For each of those candidates, we extracted its features according to subsection 3.7.2.

Table 4.6: Results of the statistical filter

	All data	Frequency filter	Frequency + Distance filter
Total candidates	157075	26484	4570
Correct candidates	6290	3664	1847
Correct English OOV translations	1540	1477	1312

Table 4.7: Strategy for imbalanced classification

Machine learning algorithms	Imbalanced classification strategy
Neural network	Parameter optimization, Sampling
k-nearest neighbor	Parameter optimization, Sampling
Naïve bayes	Parameter optimization, Sampling
Lib-supported vector machine	Parameter optimization, feature weighting, Sampling

We processed the 157,075 Chinese translation candidates with our machine learning method. We employed Rapidminer [58] as our machine learning tool. 10-fold cross validation was used for training and testing the data set.

We first ran the four base machine learning algorithms without any special modification, and then we ran the base machine learning algorithms with statistical filter, feature selection, feature weight optimization, bagging, parameter optimization, and sampling. There were 157,075 Chinese candidates, but only 6,290 of them were correct, and 1,540 English OOV terms could be translated since some English OOV terms have more than one correct translation. After the frequency filter, 26,484 Chinese candidates were left, among which only 3,664 were correct, and 1,477 English OOV terms could be translated. After the frequency and distance filter, 4,570 Chinese candidates were left, among which only 1,847 were correct, and 1,312 English OOV terms could be translated. Detailed results of the statistical filter are shown in Table 4.6. We have applied some strategies for the unbalanced classification, and a detailed chart is shown in Table 4.7.

Table 4.8: Feature selection and weighting for each machine learning algorithm

#	Features	ANN	k-NN	Naive	LibSVM	Counts
1	$\mathrm{f(c_i)}$	1	1	0	0	2
2	$\mathrm{f(e_i)}$	1	1	1	0	3
3	$\mathrm{f(e_i, c_i)}$	1	1	0	1	3
4	$\mathrm{Ff(c_i e_i)}$	1	1	1	0	3
5	$\mathrm{Bf(e_i c_i)}$	1	1	1	0	3
6	$\mathrm{SCP(c_1 \dots c_n)}$	1	1	0	1	3
7	$\chi^2\mathrm{(e_i, c_i)}$	1	0	0	0.33	1.33
8	$\mathrm{supp(e_i \rightarrow c_i)}$	1	1	1	1	4
9	$\mathrm{conf(e_i \rightarrow c_i)}$	1	1	0	1	3
10	$\mathrm{lift(e_i \rightarrow c_i)}$	1	1	1	1	4
11	$\mathrm{conv(e_i \rightarrow c_i)}$	1	0	0	0	1
12	$\mathrm{Sn(e_i)}$	1	1	0	1	3
13	$\mathbf{\lvert C_i \rvert}$	1	0	0	1	2
14	$\lvert e_i \rvert$	1	0	0	0	1
15	$\mathrm{D}(\lvert e_i \rvert, \lvert c_i \rvert)$	1	1	0	1	3
16	$\mathrm{F\overline{d}(c_i)}$	1	1	0	1	3
17	$\mathrm{B\overline{d}(c_i)}$	1	1	0.5	1	3.5
18	$\mathrm{FB\overline{d}(c_i)}$	1	1	1	1	4
19	$\boldsymbol{\delta}(\lvert e_i \rvert, \lvert c_i \rvert)$	1	1	0	1	3
20	$\mathrm{R}_{\mathbf{(c_i)}}$	1	1	1	1	4
21	$\mathrm{FR}_{\mathbf{(c_i)}}$	1	0	0	1	2
22	$\mathrm{RD}_{\mathbf{(c_i)}}$	0	0	1	1	2
23	$CDist$	0	1	1	0	2
24	$CwDist$	0	1	0	1	2
25	$\mathrm{DIC_i}$	1	0	1	0	2
# of features used for such based machine learning algorithm		22	18	10.5	16.33	

Table 4.9: Parameter optimization of base level machine learning algorithms

Machine learning algorithms	Parameter setting
Neural network	
training cycles	450
learning rate	0.3
momentum	0.01
k-nearest neighbor	
k	1
Naïve bayes	
laplace correction	on
Lib-supported vector machine	
kernel	RBF kernel
gamma	0.225595
C	0.479974
cache size	80
epsilon	0.001

Backward elimination feature selection and backward feature weight optimization were performed for four base machine learning algorithms to increase the performance. If a feature was selected by the backward elimination process for all four base machine learning algorithms, then this feature is very important to the classification of Chinese candidates. A detailed result for feature selection and feature weighting is shown in Table 4.8. Evolutionary parameter optimization was applied to each base level machine learning algorithm. A detailed parameter setting obtained from the evolutionary parameter optimization is shown in Table 4.9.

Table 4.8 shows that features such as $\boldsymbol{supp}(\boldsymbol{c_i})$, $\boldsymbol{lift}(\boldsymbol{e_i} \rightarrow \boldsymbol{c_i})$, $\boldsymbol{B\overline{d}}(\boldsymbol{c_i})$, $\boldsymbol{FB\overline{d}}(\boldsymbol{c_i})$ and $\boldsymbol{R_i}$ are considered to be very important to the classification for Chinese candidates, because they have been all selected by four different base level machine learning algorithms. The feature $\boldsymbol{conv}(\boldsymbol{e_i} \rightarrow \boldsymbol{c_i})$ is less important to the classification for Chinese candidates, because it has been selected by only one base level machine learning algorithm. The count column on the right indicates the importance of a feature. The higher the count numbers, the more important the feature is.

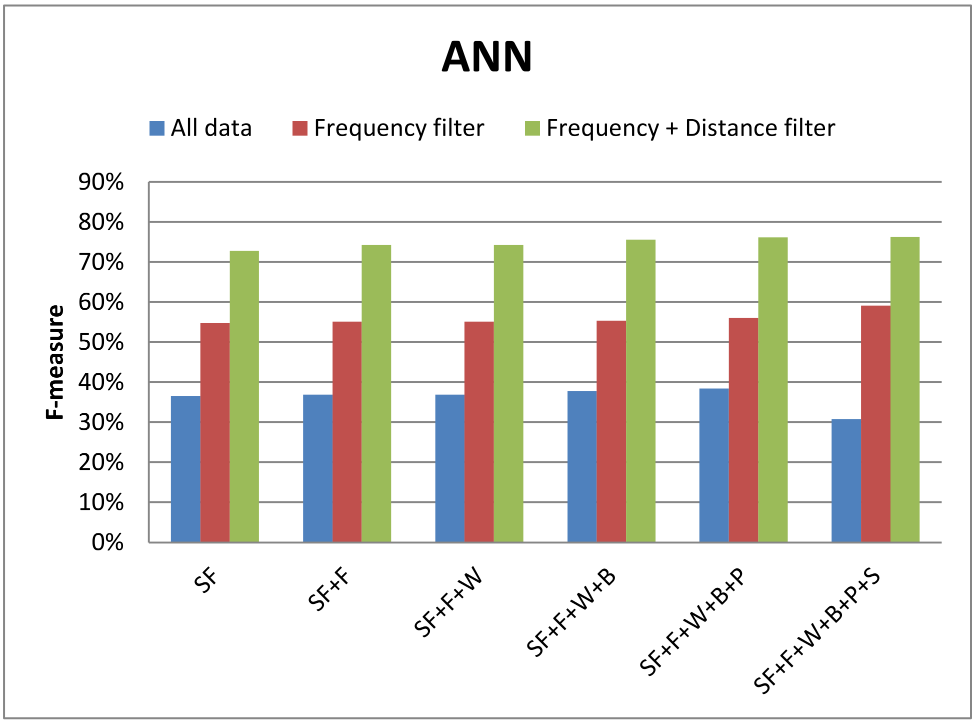


Figure 4.2: Experimental results for ANN

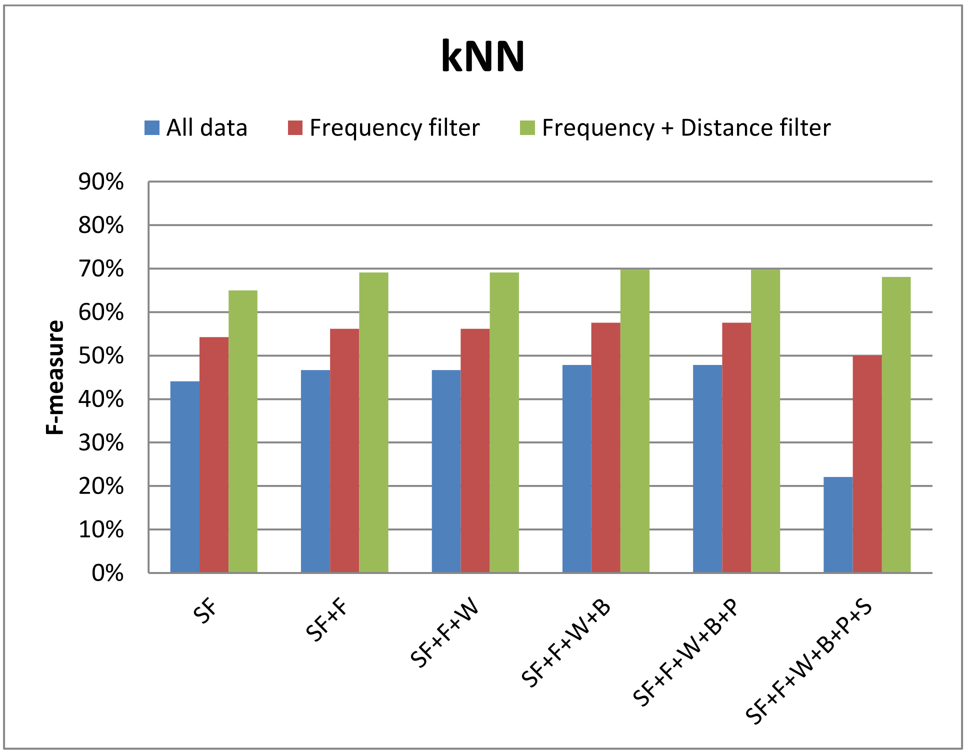


Figure 4.3: Experimental results for Knn

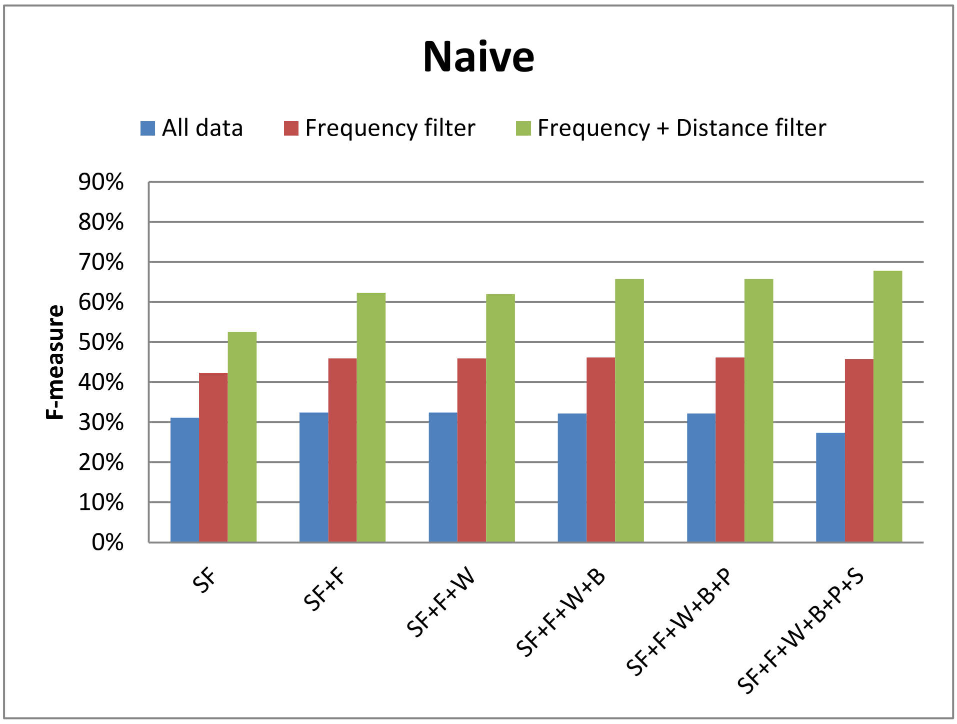


Figure 4.4: Experimental results for Naïve bayes

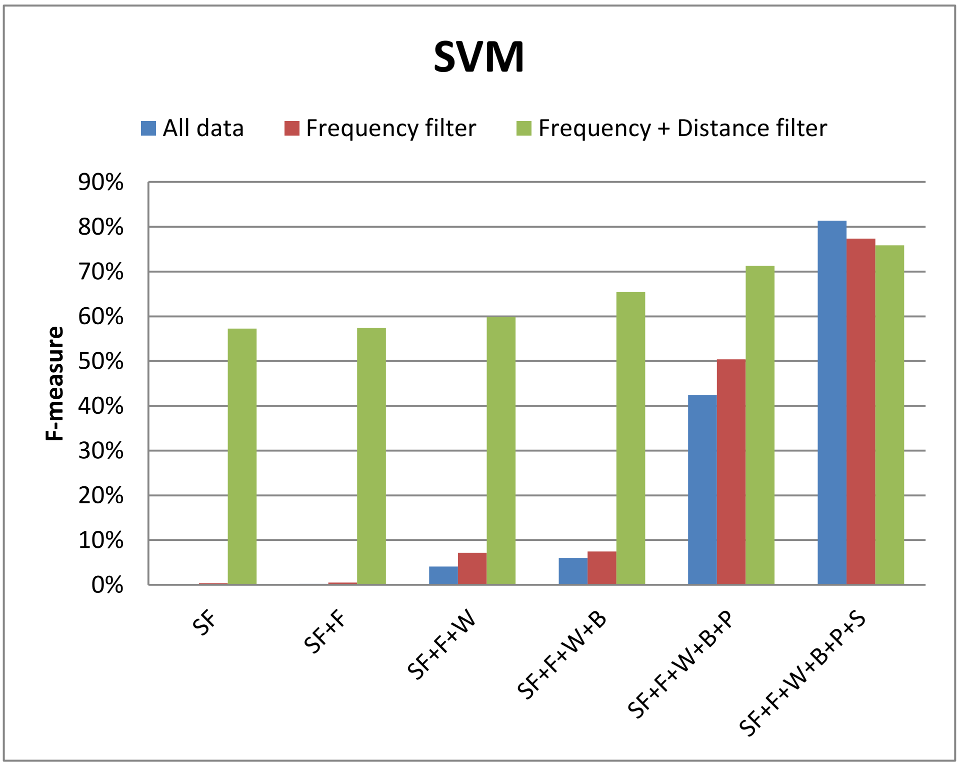


Figure 4.5: Experimental results for SVM

The experimental results are shown in Table 4.10 and Figure 4.2-Figure 4.5. The overall best performing method is the Lib-supported vector machine with statistical filter, feature selection, parameter optimization, bagging, feature weight optimization and sampling. The precision is 79.72% and the recall is 83.05%. Feature selection was not effective for Lib-supported vector machine, but feature weighting was useful for Lib-supported vector machine. Bagging was of little use, and parameter optimization was the most important for Lib-supported vector machine. Sampling made Lib-supported vector machine the only machine learning algorithm achieving high performance on data without filter (157,075 candidates). This also proves that our statistical filter is easy to use in retrieving the optimized settings. Neural network achieved the 2nd best performance with a precision of 79.01% and a recall of 73.78%, and all optimizations except feature weight optimization were useful for neural network. K-nearest neighbor gained the 3rd best performance with a precision and recall of 69.71% and 69.79% respectively, but many optimizations was not useful for k-nearest neighbor, such as feature weight optimization, parameter optimization, and sampling. The 4th performance machine learning algorithm was Naïve bayes, which achieved a precision and recall of 66.94% and 68.75% respectively. For Naïve bayes, all optimizations were useful except parameter optimization.

In order to evaluate our OOV term translation method against the existing methods, we applied the same corpus to the Length & Co-occurrence method developed by Zhang and Vines [86], the local maximum and SVM method established by Zhang, *et al*. [88], and the word/sub-word method proposed by Zhang and Sumita [81]. The length & Co-occurrence method performed a recall of only 59.56% and a precision of only 61.04%. The local maximum and SVM method was able to achieve a recall of 60.25% and precision of 67.23% (using their feature sets). The word/sub-word method gained a recall of 24.28% and precision of 30.85%. But, our method outperforms the existing methods with a recall of 83.05% and precision of 79.72%. Figure 4.6 shows the detailed results.

Table 4.10: Experimental results for each ML setup

Machine learning algorithm		All data			Data after frequency filter			Data after freauency + Distance filter		
		precision	**recall**	**f-measure**	**precision**	**recall**	**f-measure**	**precision**	**recall**	**f-measure**
SF	ANN	71.34%	24.59%	36.57%	70.68%	44.68%	54.75%	78.95%	67.64%	72.86%
SF+F		71.51%	24.86%	36.89%	70.52%	45.32%	55.18%	75.50%	73.14%	74.30%
SF+F+W		71.51%	24.86%	36.89%	70.52%	45.32%	55.18%	75.50%	73.14%	74.30%
SF+F+W+B		71.59%	25.69%	37.81%	70.97%	45.38%	55.36%	80.67%	71.23%	75.66%
SF+F+W+B+P		**72.79%**	**26.09%**	**38.41%**	71.26%	46.30%	56.13%	81.84%	71.25%	76.18%
SF+F+W+B+P+S		19.31%	75.39%	30.74%	**47.08%**	**79.44%**	**59.12%**	**79.01%**	**73.78%**	**76.31%**
SF	K-nn	44.87%	43.34%	44.09%	55.03%	53.44%	54.22%	65.42%	64.65%	65.03%
SF+F		46.92%	46.36%	46.64%	55.49%	56.82%	56.15%	69.00%	69.30%	69.15%
SF+F+W		46.92%	46.36%	46.64%	55.49%	56.82%	56.15%	69.00%	69.30%	69.15%
SF+F+W+B		**48.50%**	**47.15%**	**47.82%**	**58.18%**	**56.96%**	**57.56%**	**69.71%**	**69.79%**	**69.75%**
SF+F+W+B+P		48.50%	47.15%	47.82%	58.18%	56.96%	57.56%	69.71%	69.79%	69.75%
SF+F+W+B+P+S		12.74%	83.72%	22.11%	35.89%	82.12%	49.95%	69.02%	67.19%	68.09%
SF	Naïve	23.98%	44.56%	31.18%	36.93%	49.59%	42.33%	65.55%	43.86%	52.56%
SF+F		25.79%	43.60%	32.41%	42.72%	49.63%	45.92%	67.80%	57.72%	62.35%
SF+F+W		25.84%	43.60%	32.45%	42.52%	49.93%	45.93%	67.77%	57.20%	62.04%
SF+F+W+B		**25.26%**	**44.43%**	**32.21%**	**43.99%**	**48.68%**	**46.22%**	69.44%	62.57%	65.83%
SF+F+W+B+P		25.26%	44.43%	32.21%	43.99%	48.68%	46.22%	69.44%	62.57%	65.83%
SF+F+W+B+P+S		17.97%	57.85%	27.42%	33.25%	73.34%	45.76%	**66.94%**	**68.75%**	**67.84%**
SF	SVM	50.00%	0.08%	0.16%	17.50%	0.19%	0.38%	57.59%	56.90%	57.24%
SF+F		66.67%	0.10%	0.19%	22.50%	0.25%	0.49%	57.71%	57.12%	57.41%
SF+F+W		67.67%	2.10%	4.06%	23.50%	4.25%	7.19%	60.71%	59.12%	59.91%
SF+F+W+B		80.00%	3.13%	6.02%	28.21%	4.30%	7.46%	70.74%	60.86%	65.42%
SF+F+W+B+P		71.80%	30.13%	42.44%	74.92%	37.96%	50.39%	64.70%	79.30%	71.26%
SF+F+W+B+P+S		**79.72%**	**83.05%**	**81.35%**	**79.02%**	**75.77%**	**77.36%**	**78.74%**	**73.21%**	**75.87%**

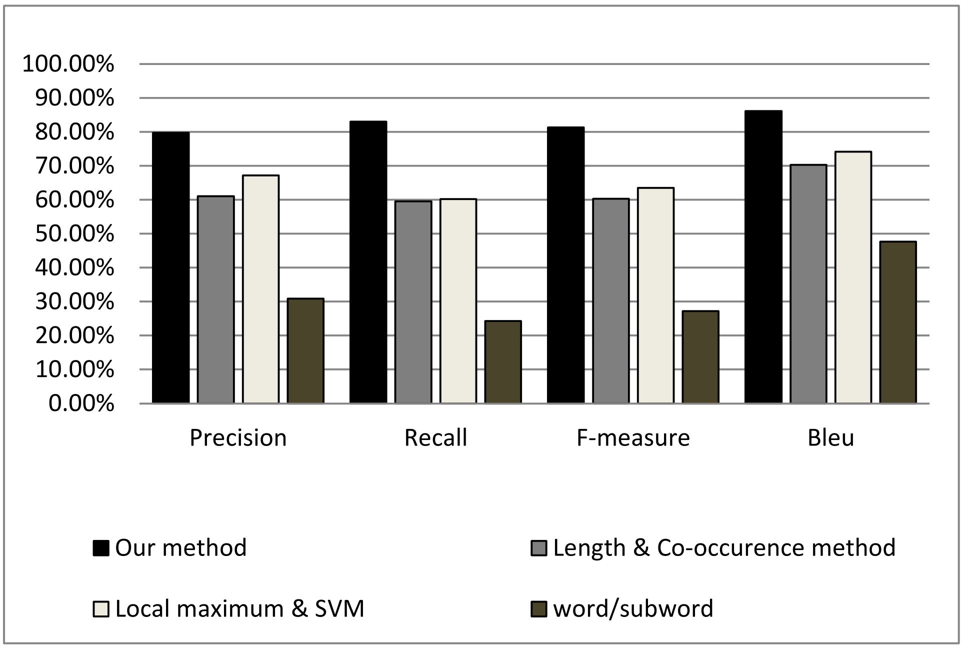


Figure 4.6: Our method compared with existing methods

However, existing methods have high performances in Bleu score, mostly because Bleu score considers the performance in n-gram. In term translation, we consider each OOV as word/sub-word. When an OOV term is half translated, Bleu score gains high precision. For example, OOV term "DiGeorge's syndrome" with the translation "症候群", Bleu score is 1. This indicates that a method can translate half of the OOV term perfectly, but it is only poorly applicable to CLIR. If "症候群" is used as the translation, many diseases unrelated to "DiGeorge's syndrome" will be retrieved by CLIR.

4.7 Experimental setup and results for CLIR of Chinese definition extraction

By using the obtained Chinese translations from section 4.5 and 4.6, we applied these translations to our method in subsection 3.9 to obtain the Chinese definitions. After querying the Chinese translations to the web, we retrieved a total of 148,816 Chinese snippets. We processed those Chinese snippets to our Chinese ranking method. Our method is

Table 4.11: Experimental results for multilingual definition extraction

	Chinese definition (accuracy)
All OOV(2,498)	1,686(67.49%)
Name type OOV(746)	635(85.12%)
Biomedical type OOV(1,752)	1051(60.00%)

Table 4.12: Experimental results for multilingual definition extraction without web features

	Chinese definition (accuracy)
All OOV(2,498)	1,394(55.80%)
Name type OOV(746)	532(71.31%)
Biomedical type OOV(1,752)	862(49.20%)

able to retrieve Chinese definitions for 1,686 OOV terms. 635 of them are name type OOV terms, and 1051 of them are biomedical type OOV terms. The detailed results are shown in Table 4.11. Furthermore, we also tested the multilingual definition extraction without web features, such as search engine ranking, domain ranking, and wiki terms ranking, the results are shown in Table 4.12.

4.8 Experimental setup and results for auto re-evaluation

Finally, we use the extracted Chinese definitions and Chinese context information to re-evaluate the correctness of Chinese translations and Chinese definitions. After comparing with human checked results, our auto re-evaluation method achieved high recalls but low precisions for translation and definition auto re-evaluation. The detailed results are shown in Table 4.13.

Table 4.13: Experimental results for auto re-evaluation

	precision	recall
All OOV translation evaluation (2,498)	46.99%	99.37%
Name type OOV translation evaluation (746)	65.25%	98.76%
Biomedical type OOV translation evaluation (1,752)	38.62%	99.84%
All OOV definition evaluation (2,498)	67.90%	99.41%
Name type OOV definition evaluation (746)	77.01%	98.78%
Biomedical type OOV definition evaluation(1,752)	62.60%	99.85%

Chapter 5

Discussions

In this chapter we provide explanations and discussions of behaviors of results obtained from methods and experiments presented in the previous chapters. In addition, we explain why our approaches perform better than existing methods and what the limitations of our approaches are. This chapter is organized as follows: In subsection 5.1, we discuss the English definition and multilingual context information extraction; In subsection 5.2, we discuss the OOV type prediction; In subsection 5.3, we discuss the name type OOV translation; In subsection 5.4, we discuss the biomedical type OOV translation; In subsection 5.5, we discuss the CLIR for Chinese definition extraction; and finally in subsection 5.6, we discuss the auto re-evaluation.

5.1 Discussion for English definition and multilingual context information extraction

Our method extracts the English definitions and context information for the OOV terms. We achieved an accuracy of 84.15% for context information extraction and an accuracy of 75.46% for English definition extraction.

Context information extraction achieved high results for both English and Chinese languages, since only English context information are extracted, Chinese context information are just dictionary translations from English context information. Our method gained high performance for English definition extraction because we considered not only the definition identification verbs, but we also considered the co-occurrences and locations of the definition identification verbs. Moreover, we considered not only the summary of the snippet but we also take the title of the snippet into consideration. Furthermore, unlike most existing methods, we utilized many web features, which contributed a lot for our high performance. Search engine ranking, domain ranking, and wiki terms ranking are very important in our method, as we can see in Table 4.2, without these features, the performance would drop dramatically. As for context information extraction, we gained high performance because we utilized many useful hyponyms of super node noun terms from WordNet, however, we only used the brief hyponyms not the full hyponyms, which may in return caused some errors for context information extraction. However, some drawbacks are within our English definition ranking method. If search engine, our algorithm, domain ranking and wiki ranking all ranked a wrong snippet highly relevant to an OOV term, we may get an error. For example, query term #128 "Fucosidosis"(a biomedical OOV term), however the snippet with high rank for "Fucosisdosis" contains definition for a person named "Fucosisdosis". The limitations of our English definition extraction method are mostly caused by OOV term that does not have a correct definition on the web, or disambiguation OOV terms. For context information extraction, it suffers from mistaken English definitions.

The Context information extraction method is a general method, thus it can be easily applied to other languages. The English definition extraction method utilize many web features and some English definition identification verbs, thus this method may be able to applied for languages that share similar characteristics with English language, such as French.

5.2 Discussion for OOV type prediction

Our method performed better because we considered the types of OOV terms. Different methods were used for different types of OOV terms. We achieved precisions of 79.76% for name type OOV term prediction and 99.93% for biomedical type OOV term prediction. We also achieved recalls of 99.86% for name type OOV term prediction and 89.21% for biomedical type OOV term prediction.

The prediction works very well for name type OOV terms, because most name type OOV terms were able to extract correct English definitions, thus the type prediction method does not suffer much from English definition errors. However a large number of biomedical OOV terms were not able to extract any useful definitions on the Internet, thus the retrieved snippets of these biomedical OOV terms contain definitions not related to biomedical. Most of those OOV terms are either terms that contain no definitions on the web or terms that are disambiguate, thus the correct definition is embedded in more detail web pages. In results, these biomedical OOV terms were wrongly classified into name type OOV terms.

The OOV type prediction method is a general method, thus it can be easily applied to other languages.

5.3 Discussion for name type OOV term translation

We achieved accuracies of 98.39% and 98.39% for translation generation and translation selection respectively. The existing methods length & co-occurrence performed accuracies of 98.39% and 93.96% for translation generation and translation selection respectively, the method of local maximum and SVM achieved accuracies of 97.72% and 95.04% for translation generation and translation selection respectively.

Many existing methods have already achieved high performances on name type OOV term translation, our method gained little improvements by using a novel statistical filter, which was able to remove some wrong translation candidates.

This proves that our filter can be used together with existing methods and increase performances.

5.4 Discussion for biomedical type OOV term translation

For biomedical OOV terms, we have evaluated our proposed methods against the existing methods on 1,752 ICD9 English biomedical OOV terms. We tested 24 combinations of machine learning approaches with 25 features on 157,075 candidates. The existing methods length & co-occurrence performed a recall of 59.56% and precision of 61.04%, the method of local maximum and SVM achieved a recall of 60.25% and precision of 67.23%, and the word/sub-word gained a recall of 24.28% and precision of 30.85%. On the contrary, our method of SF+F+W+S+B+P with the base machine learning algorithm Lib-supported vector machine surpasses the existing methods with a recall of 83.05% and precision of 79.72%.

The improvements are mainly due to the fact that our method can translate some OOV terms which the existing methods cannot handle. We build a novel self-modifying method that takes consider of each OOV term and produces different patterns to match the correct translation. Moreover, we extracted a total of 25 features, among which 11 features are novel and presented by us in this work. Some of our novel features contributed a lot for machine learning, such as modified association measures and back average distances, details of feature evaluation is shown in Table 4.8. Furthermore, our machine learning method uses a novel statistical filter system, which easily reduces the size of the data set, thus making many high cost machine learning optimization possible. Finally, optimization settings obtained from the filtered data are effectively applied back to the original data, and then high performances are achieved. However, our filter system may at the same time cause problems for Lib-supported vector machine to suffer in the unbalanced classification. The parameter optimization is the best strategy we use to solve the unbalanced classification problem in Lib-supported vector machine, but it can perform only in the filtered data due to large computational cost.

The machine learning task were experimented on a i7 PC with 48 GB of memories, yet many machine learning tasks are not possible without our statistical filter, such as parameter optimization, features selection, etc. We believe our filter many help in many other machine learning tasks where large memories are needed for computation. Moreover, we believe our self-modifying pattern construction method can be easily applied to other machine translation tasks.

5.5 Discussion for CLIR of Chinese definition extraction

We achieved accuracies of 67.49% for CLIR of Chinese definition extraction, 85.12% for name type OOV term Chinese definition extraction and 60.00% for biomedical type OOV term Chinese definition extraction.

CLIR for Chinese definition extraction achieved high results because we considered a large number of definition identification verbs. Moreover, we consider the Chinese segmentation, which made us able to extract the correct Chinese definitions even if the documents are associated to a slightly different translation, for example, "Thomas Edison" our translation "托马斯•爱迪生" document with correct definition "爱迪生". Furthermore, we considered co-occurrences and locations of definition identification verbs and OOV terms in both summary and title of the snippet. We also utilized many web features, which contributed a lot for our high performance. Search engine ranking, domain ranking, and wiki terms ranking are very important in our method, as we can see in Table 4.12, without these features, the performance would drop dramatically. Moreover, we found out that for domain ranking, ".com" is very important for Chinese language, because the largest Chinese encyclopedia Baidu uses ".com" not ".org". However, CLIR for Chinese definition extraction suffers from OOV term translation errors. If an OOV translation is wrong, then it is likely to extract a wrong definition. Moreover, CLIR for Chinese definition extraction also suffers from the source of OOV terms, especially, personal names, some source OOV terms are English famous peoples, thus definitions of these people are much easier to be found in English language than Chinese language. Finally, some biomedical OOV term does not have any Chinese definition available on the Internet.

CLIR for Chinese definition extraction method utilize many features that are suitable for languages use Kanji characters, moreover this method also utilize word segmentation, thus this method may be able to apply for other languages, which use Kanji characters and do not have word segmentation, such as Japanese.

5.6 Discussion for auto re-evaluation

We achieved a precision of 46.99% and a recall of 99.37% for translation auto re-evaluation. We also achieved a precision of 67.90% and a recall of 99.41% for definition auto re-evaluation.

We gained a high recall due to utilizing many context terms, we compared the context terms in Chinese definitions against the context terms in English definitions. The re-evaluation method suffers from the ontology trees we build in subsection 3.6, especially ontology trees for biomedical terms. Many specific context words for disease are not included in brief hyponyms of the WordNet. Furthermore, when translating these hyponyms from English to Chinese using bilingual dictionary, only first translation was used, which may result not perfectly matching terms in Chinese disease definitions.

The auto re-evaluation method is a general method, thus can be easily applied to other languages.

Chapter 6

Conclusion and future works

OOV term translation plays an important role in natural language processing, especially cross language information retrieval. In the past 20 years, many existing methods had explored various ways of finding translations for name type OOV terms in other languages. Yet, none of the existing methods offer detail information and knowledge of OOV terms for the users. The obtained translations are also OOV terms just in other language. Moreover, existing methods are not able to handle different types of OOV terms, especially biomedical type OOV terms. Most importantly, non-existing method does cross language definition extraction for OOV terms. Furthermore, it has always been very difficult for researcher and users to evaluate the correctness of translations and definitions of OOV terms. We proposed an English definition ranking method for identify the types of OOV terms and extract the monolingual definitions and multilingual context information of OOV terms. Different methods were used for translating different types of OOV terms. We have proposed a novel adaptive rules candidate extraction method for biomedical type OOV terms. We have also proposed a new statistic filter, and tested 24 different machine learning translation selection methods, which are combinations of SF, SF+F, SF+F+W, SF+F+W+S, SF+F+W+S+B, and SF+F+W+S+B+P with four base machine learning learners. Moreover, we proposed a novel CLIR for Chinese definition ranking method for extracting Chinese definitions of English OOV terms. Furthermore, we proposed an auto re-evaluate method for evaluating the correctness of Chinese translations and Chinese definitions.

We tested our methods with both name type and biomedical type OOV terms. We retrieved and processed a total of 743,914 documents (snippets). Our method achieved accuracies of 84.15% for multilingual context information extraction and 75.46% for English definition extraction respectively. Our method also achieved precision of 79.76% and high recall of 99.86% for name type OOV term prediction, and we achieved high precision of 99.93% and recall of 89.21% for biomedical type OOV term prediction. For name type OOV term translation, our method gained little improvements over existing methods with high accuracies of 98.39% and 98.39% in candidate generation and candidate selection respectively. For biomedical type OOV term translation, we have evaluated our proposed methods against the existing methods on 1,752 ICD9 English biomedical OOV terms. The existing methods length & co-occurrence performed a recall of 59.56% and precision of 61.04%, the method of local maximum and SVM achieved a recall of 60.25% and precision of 67.23%, and the word/sub-word gained a recall of 24.28% and precision of 30.85%. On the contrary, our method of SF+F+W+S+B+P with the base machine learning algorithm Lib-supported vector machine surpasses the existing methods with a recall of 83.05% and precision of 79.72%. Furthermore, our method achieved accuracies of 67.49% for Chinese definition extraction, 85.12% for name type OOV term Chinese definition extraction and 60.00% for biomedical type OOV term Chinese definition extraction. We achieved a precision of 46.99% and a recall of 99.37% for translation auto re-evaluation. We also achieved a precision of 67.90% and a recall of 99.41% for definition auto re-evaluation. In our feature work, we plan to improve more on our CLIR for Chinese definition ranking methods and auto re-evaluation method, and we plan to expend our works in CLIR for knowledge extraction.

References

1. Agrawal, R., T. Imieli, and A. Swami, *Mining association rules between sets of items in large databases.* SIGMOD Rec., 1993. **22**(2): p. 207-216.

2. Agrawal, R., T. Imielinski, and A. Swami, *Mining association rules between sets of items in large databases.* SIGMOD Rec., 1993. **22**(2): p. 207-216.

3. Agrawal, R., T. Imielinski, and A. Swami, *Database Mining: A Performance Perspective.* IEEE Transactions on Knowledge and Data Engineering, 1995. **5**(Special Issue on Learning and Discovery in Knowledge-Based Databases): p. 914–925.

4. Anderson, J.A., *An Introduction to Neural Networks*. 1995: A Bradford Book.

5. Back, T. and H.P. Schwefel, *An Overview of Evolutionary Algorithms for Parameter Optimization.* MIT Press Journals, 1993. **1**(1): p. 1-23.

6. Ballesteros, L. and B. Croft, *Dictionary-based Methods for Cross-Lingual Information Retrieval*, in *International Conference on Database and Expert Systems Applications*. 1996. p. 791-801.

7. Bergroth, L., H. Hakonen, and T. Raita, *A Survey of Longest Common Subsequence Algorithms*, in *Proceedings of the Seventh International Symposium on String Processing Information Retrieval (SPIRE'00)*. 2000, IEEE Computer Society. p. 39.

8. Bing. *Bing Search API.* 2012; Available from: http://www.bing.com/toolbox/bingdeveloper/.

9. Bishop, C.M., *Neural Networks for Pattern Recognition*. 1995: Oxford University Press.

10. Breiman, L., *Bagging Predictors.* Machine Learning, 1996. **24**(2): p. 123-140.

11. Bremner, D., et al., *Output-Sensitive Algorithms for Computing*

Nearest-Neighbour Decision Boundaries. Discrete Comput. Geom., 2005. **33**(4): p. 593-604.

12. Buitelaar, P., D. Olejnik, and M. Sintek, *A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis*, in *The Semantic Web: Research and Applications, First European Semantic Web Symposium*. 2004: Heraklion, Crete, Greece,. p. 31-44.

13. 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*. 2004, ACM: Sheffield, United Kingdom. p. 146-153.

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

15. Chien, L.-F., *PAT-tree-based keyword extraction for Chinese information retrieval.* SIGIR Forum, 1997. **31**(SI): p. 50-58.

16. Chunxia, Z. and J. Peng. *Automatic extraction of definitions*. in *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*. 2009.

17. CNN. *Fortune 500 companies 2011*. 2011; Available from: http://money.cnn.com/magazines/fortune/fortune500/2011/index.html.

18. Cormen, T.H., et al., *Chapter 17 "Greedy Algorithms"*, in *Introduction to Algorithms*. 1990, MIT Press.

19. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine Learning, 1995. **20**(3): p. 273-297.

20. Cover, T. and P. Hart, *Nearest neighbor pattern classification.* Information Theory, IEEE Transactions on, 1967. **13**(1): p. 21-27.

21. Dan, G., *Algorithms on Strings, Trees and Sequences:*. Computer Science and Computational Biology. 1997: Cambridge University Press.

22. Fellbaum, C., *WordNet An Electronic Lexical Database*. 1998.

23. Frank, E. and R.R. Bouckaert, *Naive bayes for text classification with unbalanced classes*, in *Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases*. 2006, Springer-Verlag: Berlin, Germany. p. 503-510.

24. Fung, P. and L.Y. Yee, *An IR approach for translating new words from nonparallel, comparable texts*, in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*. 1998, Association for Computational Linguistics: Montreal, Quebec, Canada. p. 414-420.

25. Government, C. *Chinese language*. 2009; Available from: http://www.china-language.gov.cn.

26. Grefenstette, G., *Cross-Language Information Retrieval*. 1998: Kluwer Academic Publishers. 182.

27. Groves, R.M., et al., *Survey Methodology*. 2004.

28. Gu, T., et al., *An ontology-based context model in intelligent environments*, in *In Communication Networks and Distributed Systems Modeling and Simulation Conference*. 2004. p. 270-275.

29. Han, J. and M. Kamber, *Data Mining: Concepts and Techniques*. 2 ed. 2006. 770.

30. Hand, D., H. Mannila, and P. Smyth, *Principles of Data Mining*. 2001, Massachusetts London: A Bradford Book The MIT Press Cambridge.

31. Hany, H. and S. Jeffrey. *An Integrated Approach for Arabic-English Named Entity Translation*. in *ACL workshop on Computational Approaches to Semitic Languages*. 2008.

32. Hirschberg, D.S., *Algorithms for the Longest Common Subsequence Problem.* J. ACM, 1977. **24**(4): p. 664-675.

33. ICD9, C. *International Classification of Diseases, Ninth Revision (ICD-9).* 2009; Available from: http://www.cdc.gov/nchs/icd/icd9.htm.

34. ICD9, C. *公告罕见疾病名單暨ICD-9-CM編碼一覽表.* 2009 30 SEP 2009;

Available from: http://web.cdc.gov.tw/public/Attachment/981111144376.pdf.

35. Jang, M.-G., S.H. Myaeng, and S.Y. Park, *Using mutual information to resolve query translation ambiguities and query term weighting*, in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. 1999, Association for Computational Linguistics: College Park, Maryland. p. 223-229.

36. Jin, B. and Z. Yan-Qing. *Support vector machines with evolutionary feature weights optimization for biomedical data classification*. in *Fuzzy Information Processing Society, 2005. NAFIPS 2005. Annual Meeting of the North American*. 2005.

37. Jinshan. *Kingsoft Translation software*. 2009; Available from: www.kingsoft.com.

38. Kang, I.-H. and G. Kim, *English-to-Korean transliteration using multiple unbounded overlapping phoneme chunks*, in *Proceedings of the 18th conference on Computational linguistics - Volume 1*. 2000, Association for Computational Linguistics: Germany. p. 418-424.

39. Kearns, M. and Y. Mansour, *On the boosting ability of top-down decision tree learning algorithms*, in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*. 1996, ACM: Philadelphia, Pennsylvania, United States. p. 459-468.

40. Kilgarriff, A. and G. Grefenstette, *Introduction to the Special Issue on the Web as Corpus*. Computational Linguistics, 2003. **29**(3).

41. Lallich, S., B. Vaillant, and P. Lenca, *A Probabilistic Framework Towards the Parameterization of Association Rule Interestingness Measures* Methodology and Computing in Applied Probability 2007. **9**(3): p. 447-463.

42. Language, C. *Languages of China*. 2009; Available from: http://www.chineselanguage.org.

43. LDC. *English/Chinese dictionary*. 2008; Available from: http://www.ldc.upenn.edu.

44. Lin, C.-Y., *Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics*, in *In **42nd ACL***. 2004, ACL. p. Article No. 605.

45. Lu, C., Y. Xu, and S. Geva, *Translation disambiguation in web-based translation extraction for English-Chinese CLIR*, in *ACM symposium on Applied computing*. 2007, ACM: Seoul, Korea. p. 819-823.

46. Lu, W.-H., L.-F. Chien, and H.-J. Lee, *Anchor text mining for translation of Web queries: A transitive translation approach.* ACM Trans. Inf. Syst., 2004. **22**(2): p. 242-269.

47. Magerman, D.M., *Statistical decision-tree models for parsing*, in *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. 1995, Association for Computational Linguistics: Cambridge, Massachusetts. p. 276-283.

48. Malaise, V. and P. Zweigenbaum. *Detecting semantic relations between terms in definitions*. in *COLING CompuTerm 2004: 3rd INternational Workshop on COmutational Terminology*. 2004.

49. Manning, C.D., P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. 2008, Cambridge University Press p. 253-287.

50. Ngomo, A.-C.N., *Knowledge-free discovery of domain-specific multiword units*, in *Proceedings of the 2008 ACM symposium on Applied computing*. 2008, ACM: Fortaleza, Ceara, Brazil. p. 1561-1565.

51. Nie, J.-Y., et al., *Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web*, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999, ACM: Berkeley, California, United States. p. 74-81.

52. O'Brien, C. and C. Vogel, *Spam filters: bayes vs. chi-squared; letters vs. words*, in *Proceedings of the 1st international symposium on Information and communication technologies*. 2003, Trinity College Dublin: Dublin, Ireland. p. 291-296.

53. Oxford. *Oxford language dictionary*. 2012; Available from: http://www.oxfordlanguagedictionaries.com/Public/PublicHome.html/.

54. Pearson, J. *The expression of definitions in specialised texts: a corpus-based analysis.* in *Proceedings of the seventh Euralex International Congress*. 1996.

55. Pei, J., J. Han, and L.V.S. Lakshmanan, *Mining Frequent Itemsets with Convertible Constraints*, in *Proceedings of the 17th International Conference on Data Engineering*. 2001, IEEE Computer Society. p. 433.

56. Pirkola, A., *The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval*, in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998, ACM: Melbourne, Australia. p. 55-63.

57. Press, W.H., et al., *Section 16.5. Support Vector Machines*, in *Numerical Recipes 3rd Edition:The Art of Scientific Computing*. 2007, Cambridge University Press.

58. Rapidminer, *Rapidminer data mining tool.* 2009.

59. Ren, F., *From Cloud Computing to Language Engineering, Affective Computing and Advanced Intelligence.* International Journal of Advanced Intelligence, 2010. **2**(1): p. 1-14.

60. Ren, F. and D.B. Bracewell, *Advanced Information Retrieval.* Electron. Notes Theor. Comput. Sci., 2009. **225**: p. 303-317.

61. Resnik, P., *Mining the Web for bilingual text*, in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. 1999, Association for Computational Linguistics: College Park, Maryland. p. 527-534.

62. Safavian, S.R. and D. Landgrebe, *A survey of decision tree classifier methodology* IEEE Transactions on Systems, Man, and Cybernetics, 1991 **21**: p. 660-674.

63. selfcreation. *Famous people of all times*. 2010; Available from: http://www.selfcreation.com/creation/famous_people.htm.

64. Shah, S. and A. Kusiak, *Cancer gene search with data-mining and genetic algorithms.* Computers in Biology and Medicine archive, 2007. **37**(2): p.

251-261.

65. Shi, L., *Mining OOV Translations from Mixed-Language Web Pages for Cross Language Information Retrieval*, in *Advances in Information Retrieval*, C. Gurrin, et al., Editors. 2010, Springer Berlin / Heidelberg. p. 471-482.

66. Sierra, G., et al. *Web Exploitation for Definition Extraction*. in *Web Congress, 2009. LA-WEB '09. Latin American*. 2009.

67. Silva, J.F.d., S.G. Jos, and G.P. Lopes, *Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units*, in *Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*. 1999, Springer-Verlag. p. 113-132.

68. Silva, J.F.d. and G.P. Lopes. *Extracting Multiword Terms from Document Collections*. in *In Proceedings of the VExTAL,Venezia per il Trattamento Automatico delle Lingu*. 1999. Universiá Cá Foscari, Venezia.

69. Silva, J.F.d. and G.P. Lopes, *A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora*, in *Sixth Meeting on Mathematics of Language*. 1999 University of Central Florida, Orlando, Florida, USA. p. 369-381.

70. Tichy, W.F., *The string-to-string correction problem with block moves.* ACM Trans. Comput. Syst., 1984. **2**(4): p. 309-321.

71. Tiffin, N., et al., *Integration of text and data-mining using ontologies successfully selects disease gene candidates.* Nucleic Acids Res, 2005. **33**: p. 1544–1552.

72. Udupa, R., et al., *"They Are Out There, If You Know Where to Look": Mining Transliterations of OOV Query Terms for Cross-Language Information Retrieval Advances in Information Retrieval*, in *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*. 2009, Springer Berlin / Heidelberg. p. 437-448.

73. Viriyayudhakorn, K., T. Theeramunkong, and C. Nattee, *Mining Translation pairs for Thai-English Medical Terms*, in *KICSS*. 2008, KICSS. p. 204-211.

74. WiKi, C. *Chinese language*. 2009; Available from:

http://en.wikipedia.org/wiki/Chinese_language

75. wikipedia. *List of car brands*. 2010; Available from: http://wiki.answers.com/Q/List_of_car_brands.

76. Witten, I.H. and E. Frank, *Data mining: practical machine learning tools and techniques with Java implementations.* SIGMOD Rec., 2002. **31**(1): p. 76-77.

77. Wu, Z. and G. Tseng, *Chinese text segmentation for text retrieval: Achievements and problems.* Journal of the American Society for Information Science and Technology, 1993. **44**(9): p. 532-542.

78. Yang, C.C. and K.W. Li, *Automatic construction of English/Chinese parallel corpora.* J. Am. Soc. Inf. Sci. Technol., 2003. **54**(8): p. 730-742.

79. YE, N. and Q. CHEN, *An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems.* Quality and Reliability Engineering International, 2001. **17**(2): p. 105-112.

80. Yuan, Y. and M.J. Shaw, *Induction of fuzzy decision trees.* Fuzzy Sets and Systems, 1995. **Volume 69**(Issue 2).

81. Zhang, R. and E. Sumita. *Chinese unknown word translation by subword re-segmentation*. in *3rd International Joint Conference on Natural Language Processing*. 2008. Hyderabad, India.

82. Zhang, T., *Association Rules*, in *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*. 2000, Springer-Verlag. p. 245-256.

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

84. Zhang, Y., F. Huang, and S. Vogel, *Mining translations of OOV terms from the web through cross-lingual query expansion*, in *ACM SIGIR conference on Research and development in information retrieval*. 2005, ACM: Salvador, Brazil. p. 669-670.

85. Zhang, Y. and P. Vines, *Detection and translation of OOV terms prior to query time*, in *ACM SIGIR conference on Research and development in information retrieval*. 2004, ACM: Sheffield, United Kingdom. p. 524-525.

86. Zhang, Y. and P. Vines, *Using the web for automated translation extraction in cross-language information retrieval*, in *ACM SIGIR conference on Research and development in information retrieval*. 2004, ACM: Sheffield, United Kingdom. p. 162-169.

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

88. Zhang, Y., Y. Wang, and X. Xue, *English-Chinese bi-directional OOV translation based on web mining and supervised learning*, in *ACL-IJCNLP 2009 Conference Short Papers*. 2009, Association for Computational Linguistics: Suntec, Singapore. p. 129-132.

89. Zhang, Y., Y. Wang, and X. Xue, *English-Chinese bi-directional OOV translation based on web mining and supervised learning*, in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. 2009, Association for Computational Linguistics: Suntec, Singapore. p. 129-132.

90. Zhou, D., et al., *A Hybrid Technique for English-Chinese Cross Language Information Retrieval.* ACM Trans. Comput. Syst., 2008. **7**(2): p. 1-35.

Appendix A

Curriculum Vitae

Education:

Ph.D in Information Science, Japan Advanced Institute of Science and Technology, Japan

(October 2010-Present)

M.S.IT in Information Science, Thammasat University, Thailand

(November 2007- March 2010)

B.BA in Business administration and finance, **Summa Cum Laude honor**,

Ramkhamhaeng University, Thailand

(June 2003-November 2006)

Publications:

International Journals

- Jian Qu, Thanaruk Theeramunkong, Minh Le Nguyen, Akira Shimazu, Cholwich Nattee and Pakinee Aimmanee: A flexible rule-based approach for discovering medical English-Chinese OOV term translations from web with machine learning. International Journal of Computer Processing of Oriental Languages. Accepted. To be appear in volume 24, 2013.

- Jian Qu, Thanaruk Theeramunkong, Cholwich Nattee and Pakinee Aimmanee: Web Translation of English Medical OOV terms to Chinese with Data Mining Approach. Thammasat International Journal of Science and Technology, Vol. 16, No. 2, 26-40. April-June 2011.

- Jian Qu, Minh Le Nguyen and Akira Shimazu: Cross language information extraction and auto evaluation for OOV term translations. IEEE International Journal of China Communications. Under review. Invited submission.

Lecture notes

- Jian Qu, Akira Shimazu and Minh Le Nguyen: OOV Term Translation, Context Information and Definition Extraction Based on OOV Term Type Prediction. Lecture Notes in Computer Science Volume 7614, 2012, 76-87. Springer.

International Conferences

- Jian Qu, Nguyen Le Minh, Akira Shimazu and Takaya Yuizono: Integrated ranking approach for OOV term definition and multilingual context information extraction, 8th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE'12), 56-63. HeFei, China, Sep 2012. **(Best Paper Award)**

- Jian Qu, Nguyen Le Minh and Akira Shimazu: Web based English-Chinese OOV term translation using Adaptive rules and Recursive feature selection, 25th Pacific Asia Conference on Language, Information and Computation (PACLIC'25), 1-10. Singapore, Dec 2011. Published by Waseda University.

- Jian Qu, Thanaruk Theeramunkong, Cholwich Nattee and Pakinee Aimmanee: A Novel Candidate Generation Technique for Web based English-Chinese

Medical OOV Term Translation. 4th International Conference on Knowledge, Information and Creativity Support Systems(KICSS'09), 61-68. Seoul, Korea, Nov 2009. Published by JAIST.

- Jian Qu, Kobkrit Viriyayudhakorn, Thanaruk Theeramunkong, Cholwich Nattee and Pakinee Aimmanee: Automatic English to Chinese Translation of Medical Terms using Association Rule Mining with Web Data. 6th Joint Conference on Computer Science and Software Engineering(JCSSE'6), 336-341. Phuket, Thailand, May 2009.