Sirindhorn International Institute of Technology
Thammasat University

Thesis IT-MS-2010-01

A NOVEL APPROACH ON CANDIDATE GENERATION FOR AUTOMATIC ENGLISH TO CHINESE MEDICAL TERM TRANSLATION USING DATA MINING WITH WEB DATA

Jian Qu

A NOVEL APPROACH ON CANDIDATE GENERATION FOR AUTOMATIC ENGLISH TO CHINESE MEDICAL TERM TRANSLATION USING DATA MINING WITH WEB DATA

A Thesis Presented

by

Jian Qu

Master of Science

Information technology program

Sirindhorn International Institute of Technology

Thammasat University

May 2010

A NOVEL APPROACH ON CANDIDATE GENERATION FOR AUTOMATIC ENGLISH TO CHINESE MEDICAL TERM TRANSLATION USING DATA MINING WITH WEB DATA

A Thesis Presented

By

Jian Qu

Submitted to

Sirindhorn International Institute of Technology

Thammasat University

In partial fulfillment of the requirement for the degree of

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

Approved as to style and content by

Advisor and Chairperson of Thesis Committee	Pakinee Aimmanee, Ph.D.
Co-Advisor & Committee Member	Asst. Prof. Cholwich Nattee, D.Eng.
Committee Member and Chairperson of Examination Committee	Assoc. Prof. Thanaruk Theeramunkong, D.Eng.
Committee Member	Assoc. Prof. Ekawit Nantajeerawarawat, D.Eng.

External Examiner: Prof. David Wai-lok Cheung, Ph.D.

May 2010

Acknowledgement

It has been a long time and a struggle process to achieve a research based master degree, especially in one of the best universities in Thailand. 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 Associated Professor Dr. Thanaruk Theeramankong who offered me a chance to transform from a business undergraduate to a scientific researcher, his kind support during my research made me possible of completing this master degree. Assistant Professor Dr. Cholwich Nattee guided my research, he taught me from the beginning of how to conduct research. I remember at the first I was interesting in building a huge CLIR system, but Dr. Cholwich advised me to look in detail of CLIR problems and research into more advanced topics. One of the most important people during my research is my advisor-Dr. Pakinee Aimmanee. She took me as one of her master student regardless of my business background, she thought me programming language and how to conduct research, she also offered me the NSTDA and SIIT scholarship, which covers my tuition fees, and she taught me in constructing and writing this thesis. I wish to express my deepest gratitude to Dr. Thanaruk Theeramankong, Dr. Choliwich Nattee, Dr. Pakinee Aimmanee and Dr. Ekawit Nantajeerawarawat who guide me during my research and made me possible of publishing two international conferences and one international journal. I would also like to express my sincere gratitude to the curse instructing professors during my study, Dr. Thanasak Wongtanakitcharoen who taught me the numeric method and Professor Dr. R. H. B. Excel who taught me research method knowledge.

I would like to express my sincere gratitude to my financial sponsor NSTDA and SIIT, and Miss. Waraporn Thongthua the head of student affairs who takes care of all the paper requirements and documents for the scholarship. I also like to express my deepest gratitude to my boss-Mr. Sanan Eksangkul and his company-Eason Paint Public Company Limited for offering me time and support for my study during my working hours.

During my research life, most of our KINDML lab members assisted me on my work. Xinming Zhao and Kobkrit Viriyayudhakorn were the most contributors who taught me in detail of computer language and coding. Nhat Quang Tran guided me some basics of engineering math. Jakkrit TeCho explained me many data mining theories and tools, which made my research possible. I would record my thanks to Thawatchai Suwanapong, Ithipan Methasate, Nichchanan Kittiphattanabawon and all KINDML members for their kind support during my lab life. Finally, grateful acknowledgement is extended to all ICT and IT staff for their kind support.

Abstract

A major problem of the CLIR is the Out of vocabulary (OOV) terms, which are typically new terms that cannot be found in dictionaries. It has always been so difficult to successfully translate OOV terms form one language to other, especially for those languages without clear word boundaries such as Chinese. Some Chinese translations of English medical OOV terms are not perfectly translated due to non-Chinese characters. All existing English to Chinese OOV term translation approaches assume Chinese translations only contain Chinese characters, as a result non-Chinese characters which sometimes are important in the Chinese translations are ignored. This problem usually does not occur in name entity OOV terms but it becomes problematic for technical and medical type OOV terms. We solve the non-Chinese character problem with a rule based candidate generation approach which considering different English OOV term and automatically adjust the candidate generation system accordingly. We also proposed some new features such as modified association measures and average location distance of translation candidate to improve the accuracy of translation. We used a machine learning system to select the correct Chinese translations. Our approach outperforms the existing approaches with a precision of 95.48% and can handle the Chinese translation that includes non-Chinese characters.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Today information spreads fast throughout our world in many different ways such as newspapers, televisions and the Internet. Neither newspapers nor televisions have the fast spreading speed and geographical free availability of the Internet. The Internet offers us an easy and fast access to information worldwide with no boundaries. The amount of information on the World Wide Web is increasing every day; 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 English may only understands the English web pages. Although there were many translation tools proposed, they work well only with simple words or short sentences. The Internet users need an automatic multilingual translation tool. In this chapter, we are going to describe the cross-language information retrieval and the out of vocabulary problem, our motivation, objectives, 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

Cross-language information retrieval (CLIR) 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. 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 (Jinshan 2009), 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 out of vocabulary problem.

Out of vocabulary (OOV) terms are typically new terms from current affairs, such as person 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 entities such as person names, brand names, and location names, and 2) the technical OOV terms such as newly technological terms and medical terms. Name entities are usually translated by transliteration, and, the Chinese translation usually includes only Chinese characters. Technical OOV terms are usually translated by a combination of meaning and transliteration. Then, some Chinese translations include non-Chinese characters. One way to deal with OOV term problems is to add many dictionaries from different fields to include the translations of that OOV term. Some disadvantages for this method includes: require lots of memory for dictionaries; some OOV terms are too new, even the latest dictionary does not include these OOV terms. Another way is to use the huge amount of information available on the Internet to translate the OOV terms. The translations of the OOV terms usually co-occur with the OOV term on the same document, if we can extract these translations, we could build an automatic OOV term translation system without human interfere.

1.2 Motivation

Information technology in Natural language processing field has presented number of methods on English to Chinese OOV terms translation. Chinese is a language spoken by over one billion populations in the world, and it is the second largest spoken language besides English. All existing methods for English to Chinese OOV terms translation take the advantage of the Internet. Most of them mainly focus on frequently used name entity OOV terms. Name entity OOV terms are typically famous person names, location names, brand names, which occur frequently on newspaper articles and their translations are usually demand by the general public. However, the rapid growing of cross-country patients, increasing cross-nation medical researchers, and growing multinational hospitals, also needs the translations of some special medical OOV terms. Whenever a new sickness is been discovered, usually it is initially been given a name in English. Most the time, the special medical terms are too difficult to be translated, so the original English terms are been used as Chinese translation. It is difficult for Chinese medical researchers and patients to understand the meaning of the disease if such English medical OOV term cannot be presented in Chinese language.

It has always been so difficult to successfully translate OOV terms from one language to another, especially for those languages without clear word boundaries such like Chinese. Most existing name entity OOV term translation approaches usually produce multiple answers that require human to select. It is reasonable and easy for human to differentiate a brand name and a normal term. Nevertheless, those methods are not suitable for technical OOV terms, especially medical OOV terms, because it is difficult for non-expert persons to select the correct translation as those OOV terms are usually technical terms. According to our observation, some English medical OOV terms are not perfectly translated into Chinese as they sometimes include non-Chinese characters such

as English alphabets, numbers, and symbols. At the present, the existing English to Chinese OOV term translation system ignore those non-Chinese characters in the Chinese translations.

1.3 Objectives

We propose to build a system that translates English medical OOV terms into Chinese. This system uses the Internet as the source for Chinese translations. Different from most existing approaches, our system considers the non-Chinese characters in the Chinese translations. Our system selects the Chinese translations automatically using association measures and data mining techniques.

1.4 Scope of our work

This thesis focuses on translating English Medical OOV term to Chinese only since existing English OOV term to Chinese translation approaches cannot perfectly handle English medical OOV terms. This thesis looks in detail when applying the existing English OOV term to Chinese translation approaches to the English medical OOV term problems and finds ways to improve the quality of the results in terms of accuracy. Certain OOV terms are beyond our translation ability. Our system is limited to the English OOV terms that can only be retrieved from the Chinese Web. We only consider the Chinese translations that are not too far away from the English OOV term.

1.5 Thesis Overview

The overall thesis is outlined in the following chapters. The existing English to Chinese OOV terms translation approaches are presented in Chapter 2, which is a summary of extensive literature review. In Chapter 3, we present our proposed approaches for English-Chinese OOV term translation. In Chapter 4, we present the experimental results of our proposed approaches. In Chapter 5, we present the detailed discussion of those experiment results. Finally, we conclude the thesis in Chapter 6.

Chapter 2

English to Chinese OOV term Translation Background

2.1 Introduction to Chinese

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 more than 370,000 Chinese words.(Government 2009; Language 2009; WiKi 2009)

2.2 OOV Term Translation Problem

The out-of-vocabulary are words arise from daily actives, such as person 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 (OOV) 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 and cross language translation software. For our work, we focus on the problems of OOV terms arisen in the cross language translation.

One way to do the cross language translation is to add as many as dictionaries from many fields to the system hoping that they contain the needed word. However, this way is not very efficient as we need a lot of memory to store a large size of dictionary and more importantly the needed word may be too new so that it is not yet been included in any dictionaries.

Another way without helps of any other dictionaries to get the cross language translation of OOV terms is to find their existing translations available on the internet. This approach is quite convenient but requires many additional steps to overcome language barriers such as no word boundary in some languages like Japanese, Thai, or Chinese.

2.3 Standard Procedures for OOV terms Translation using the Internet

When we focus on the web mining, there are four steps for the translations of the OOV terms. Those steps are 1) getting the text document that contains the OOV terms and the possible translations on the Internet, 2) generating possible translation candidates for OOV term, 3)filtering translation candidates; and 4) finally selecting the correct translation candidates.

In the following subsection, we review those standard processes that are used to retrieve the text document that contain the OOV terms and the possible translations on the internet.

2.3.1 Getting Web pages Containing Possible OOV Terms’ Translations

Many multi-languages 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 web pages, the Chinese translations for these English OOV terms are usually nearby (Zhang and Vines 2004). Cheng et al. observed that if a Chinese term occurs in an English web page, its translation usually exists in the same page as well (Cheng, Teng et al. 2004). A number of existing approaches have presented techniques for retrieving the translation pairs from multi-languages 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 (CLIR) (Fung and Yee 1998; Nie, Simard et al. 1999; Resnik 1999; Kilgarriff and Grefenstette 2003; Yang and Li 2003).

2.3.1.2 The anchor text based approach

Proposed by Lu, Chien, and Lee 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 links to a number of web pages in a target language that possibly contain the translation of our OOV term query (Lu, Chien et al. 2004).

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 below (Cheng, Teng et al. 2004; Lu, Chien et al. 2004; Zhang and Vines 2004).

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 (Zhang, Huang et al. 2005).

2.3.2 Generating possible Chinese translation candidates

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 webs 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.

There are two existing approaches, which are considering all substrings and using predefined patterns. The details of each method are provided in the next section.

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 (Zhang and Vines 2004). They used a candidate length and co-occurrence rule and achieved up to 80% accuracy to translate name entities of English OOV terms such as person 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 twenty 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 (Cheng, Teng et al. 2004; Zhang and Vines 2004; Lu, Xu et al. 2007).

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 (Viriyayudhakorn, Theeramunkong et al. 2008).

2.3.3 Filtering possible Chinese translation candidates

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 Ranking function

Given a list of tremendously many possible translation candidates generated by all substrings of the retrieved snippets, a ranking function is used for ranking the substrings based on the likelihood of being the correct translation. The ranking function *r* can be described as in the following formula (Zhang and Vines 2004):

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

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 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 build 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 mutual 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 (Jang, Myaeng et al. 1999; Lu, Xu et al. 2007).

2.3.3.2.1 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 (Silva and Lopes 1999). 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 (Lu, Xu et al. 2007; Ngomo 2008).

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

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 as follows:

Table 2-1 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.

2.3.3.2.2 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) (Cheng, Teng et al. 2004) 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, symmetric conditional probability (SCP) is used to measure how likely the candidate is a truly word term to a certain degree(Silva and Lopes 1999) and context dependency (CD) (Chien ; Gu, Wang et al. 2004) 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:

$$SCPCD(c_1 \dots c_n) = SCP(c_1 \dots c_n) \times CD(c_1 \dots c_n) \tag{2.3}$$

where the calculation of SCP and CD are

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

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

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

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

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 $f(c_1 \dots c_n)$ and $f(c_{i+1} \dots c_n)$ is the frequency of the front part and the back part of the target translation candidate. $LC(c_1 \dots c_n)$ and $RC(c_1 \dots c_n)$ 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 2-2.

Table 2-2 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 Selecting the correct translation candidates

As the Target-language translation candidates for the source OOV term we filter in the previous step are not always the correct OOV term translation, the final process is to automatically select the correct translation for the OOV term. While some closely related languages such as English and French, Chinese and Japanese 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 usually solve the translation candidate selection process statistically, 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 (Zhang and Vines 2004). 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 (Dan 1997; Silva, Ga et al. 1999; Lin 2004):

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

Longest common suffix takes two strings $((c_1 \dots c_p), (d_1 \dots d_q))$ 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(Hirschberg 1977; Bergroth, Hakonen et al. 2000).

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

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(Tichy 1984). Some examples of *LCSuff* are shown below.

$$LCSuff(\text{知}, \text{英}) = 0$$

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

$$= 0 + 1$$

$$= 1$$

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

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

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

$$= 0+1+1+1$$

$$=3$$

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 below.

Table 2-3 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

Table 2-3 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.

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 (Lu, Xu et al. 2007):

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

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 below:

Table 2-4 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

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

$$= 1.223$$

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(O'Brien and Vogel 2003). 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 web pages 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 (Buitelaar, Olejnik et al. 2004; Cheng, Teng et al. 2004).

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

where the meaning of each variable is explained as follows

Variable	Meaning
$Eoov$	source OOV term
$(c_1 \dots c_n)$	translation candidates

a	$S(Eoov \wedge (c_1 \dots c_n))$
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(YE and CHEN 2001). 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 2-5 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

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

$$= 23.9458$$

2.4 Related Data Mining Techniques

One of the most important new emerging knowledge in artificial intelligent is data mining or sometimes called data digging. Data mining is the process of extracting hidden patterns from the data. There are few data mining methods that can be used to find the relationships between items and classify items into groups, such data mining methods are association measures and decision tree.

2.4.1 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(Zhang 2000).

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 (Agrawal, Imieli et al. 1993; Hand, Mannila et al. 2001; Pei, Han et al. 2001).

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 (Lallich, Vaillant et al. 2007) 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 (Han and Kamber 2006).

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

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.

2.4.2 Decision tree

Decision tree can classify the items into correct and wrong groups. Following we give a brief introduction of decision tree and how it classify items (Safavian and Landgrebe 1991).

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 correct and wrong 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 (Yuan and Shaw 1995; Kearns and Mansour 1996).

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. We describe the details of two steps below (Agrawal, Imielinski et al. 1995; Magerman 1995; Witten and Frank 2002; Han and Kamber 2006).

In the first step, we need to construct a training data set from different items. For the training data set, we have to manually label two different classes: the correct class and the wrong class. With the training data set, we can construct the decision tree. The decision tree tries to construct the items into a tree-like graph. Starting with the root, feature with high desecration power divide the items into few big branches (leaves). Then features with low desecration power divide the big branches (leaves) into smaller branches (leaves). Eventually all items are classified into two groups, the correct group and the wrong group. For each of those tests, decision tree tries to choose the best feature that offers the best result. If any features that in the training data set is not used in the decision tree, then those features are either with very low desecration power or cannot be used to desecrate the items at all. In the second step, each item in the experimental data set is classified by its features according to the decision tree.

Chapter 3

Our Approaches

In this chapter we describe our approaches to find OOV term translations. We propose two different approaches namely the longest common substring candidate generation with association measure approach and rule based candidate generation with data mining approach. We describe them in more detail in section 3.1 and 3.2.

3.1 Longest common substring candidate generation with Association Measures Approach

We introduce this approach with a system design and detailed method procedures.

3.1.1 System design

The objective of this approach is to build an automatic translation system to perform the translation of English medical OOV terms to Chinese medical terms. The reason we focused on medical OOV terms is because that most existing OOV term translation systems focus on name entities (person names, location names, and brand names). When applying those methods to medical OOV terms, we encounter several problems. Existing OOV term translation approaches usually consider the co-occurrence frequencies and the length of the OOV translation candidates, which sometimes obtain partial answers of the translation. These existing approaches usually require human judgment to select correct translation from multiple translation candidates (Zhang and Vines 2004; Zhang, Vines et al. 2005).

The various English-to-Chinese OOV term translation techniques are suggested by several literatures that texts to be performed without word segmentation because their OOV terms are name entities for example person names, brand names, and location names. Since those Chinese translations of name entities do not exist in dictionary, the dictionary based word segmentation system would usually segment those OOV terms into smaller sequences of characters or individual characters (Cheng, Teng et al. 2004; Zhang and Vines 2004). For example, a brand name English OOV term “Intel”, if its translation in Chinese “英特尔” is applied to a word segmentation system the result is “英”, “特”, “尔” in individual Chinese characters.

In our approach, since we focus on medical terms, they have a special characteristic:

its sub-words normally have meanings. The combination of sub-words translation have a high chance to be the correct translation, so that using dictionary based segmentation system can help us scope out the boundaries of Chinese OOV term translations.(Wu and Tseng ; Ponte and Croft 1996; Cheng, Young et al. 1999). For example, a English medical OOV term “Urea cycle disorder”, if its translation in Chinese “尿素循環障礙” is applied to a word segmentation system the result is “尿素”(Urea), “循環”(cycle), “障礙”(disorder) in Chinese words. We describe a technique to extract the correct Chinese translation from English OOV terms by apply this approach to take consideration of distance, co-occurrence frequencies and length of the translation candidates. Among all of these, we improve the approach by applying association measures for better translation selection.

3.1.2 Procedures

The procedures are as follows:

Step 1. Extracting the snippets containing English medical OOV terms from the web

Step 2. Segmenting the web snippets obtained from Step 1.

Step 3. Extracting Chinese translation candidates of the English OOV terms

Step 4. Filtering the Chinese translation candidates

Step 5. Selecting the Chinese translation candidates using association measures

These steps are summarized in the diagram shown in Fig.3.1, the details of each step are described below:

3.1.3 Extracting Snippets containing English Medical OOV Terms

We query the English medical OOV terms to a search engine with special option setting. First, we set the output language on the search engine to be only Chinese so that the returning snippets contain English medical OOV terms in Chinese web pages. Second, we only retrieve the snippets of the full phrase search of the English medical OOV terms. In some search engines like Yahoo, or Google, putting the searching terms in a pair of double quotes (“ ”), it returns us the full phrase search because only snippets that contain full English OOV terms are interested. An example of retrieved snippet is shown in Fig. 3.2.

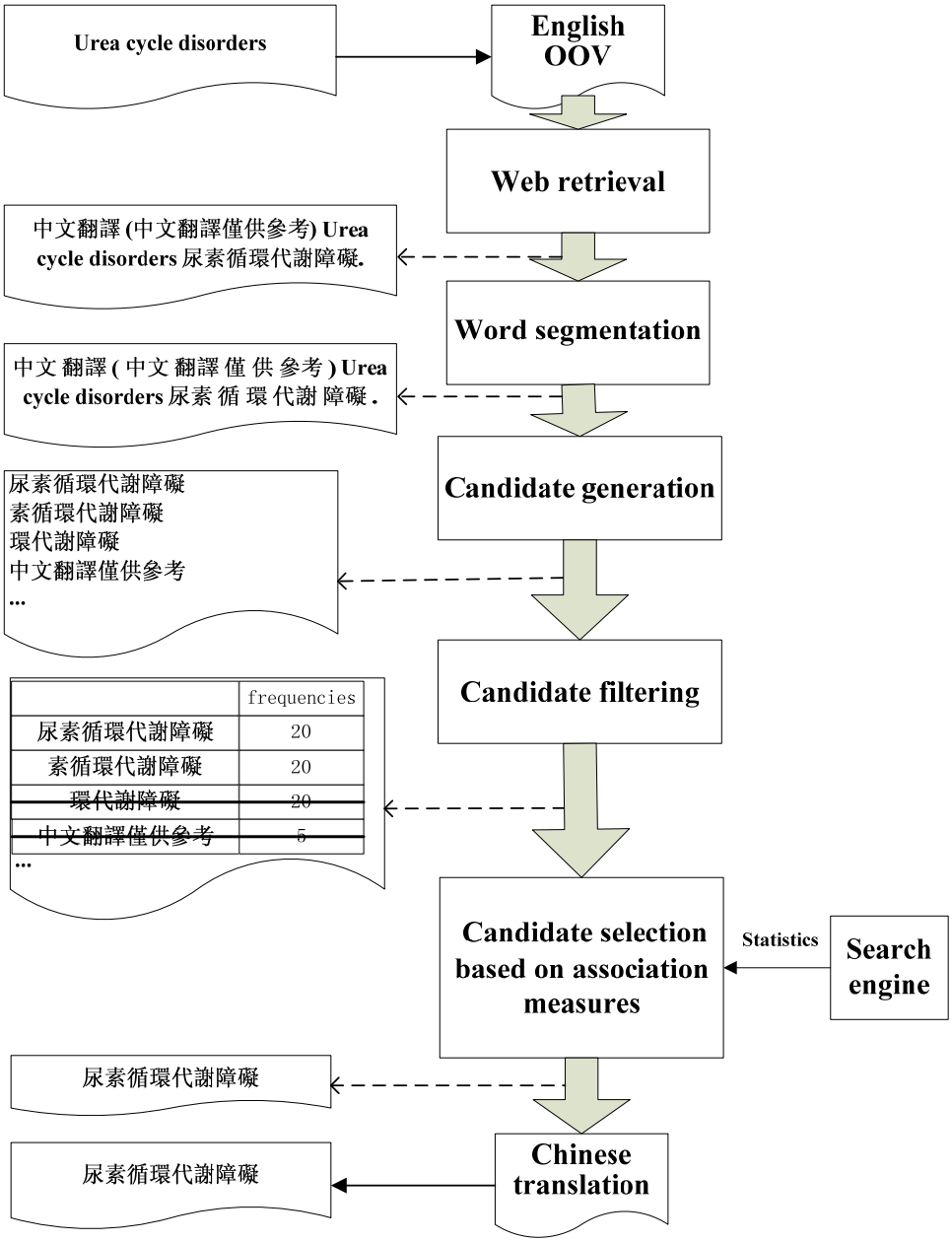

Figure 3.1 Flow chart of longest common substring candidate generation with association measure approach

-台灣醫療網 - **(Title)**

【回首頁】. 疾病名稱 中文翻譯 (中文翻譯僅供參考) Urea cycle disorders 尿素循環代謝障礙. Citrullinemia 瓜胺酸血症. Amino acid metabolic disorders 胺基酸代謝疾病 (又稱 Aminoacidopathies) Hypermethioninemia 高甲 ... **(Summary)**

http://tw16.net/monographData.asp?m1No=7&m2No=28&m3No=188&mMo=609 **(URL)**

Figure 3.2 Example of the retrieved snippet for query "Urea cycle disorders"

In the process, HTML tags are removed and each obtained web snippet is separated into three different fields into the database: URL, Title, and Summary.

3.1.4 Segmenting the web snippets

Dictionaries have been very useful during CLIR (Ballesteros and Croft 1996; Hull and Grefenstette 1996; Lu, Xu et al. 2007). In this stage, we first scan through the snippets returned from fetching queries to the web, and filter out the snippets that do not contain any Chinese characters. We run the data through a dictionary based Chinese segmentation system. Unlike other approaches that select a specific number of characters before and after the OOV term (windows size) which normally include incorrect translation candidates, in our approach we improve this window size by using the number of words resulting from word segmentation. An example of word segmentation is shown below:

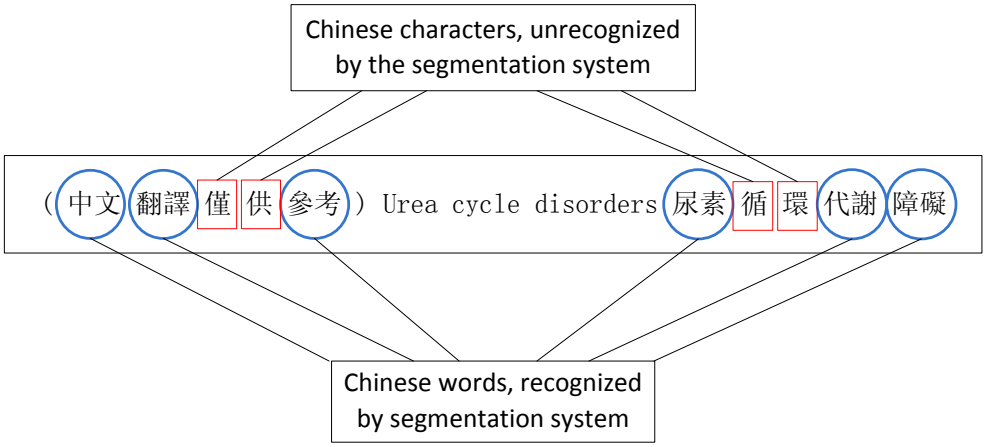

Figure 3.3 the data after segmentation

In Fig. 3.3, “Urea cycle disorders” is an English medical OOV term; “(中文翻譯僅供參考) Urea cycle disorders 尿素循環代謝障礙” is a string from snippet that is sent to the segmentation process. Since 中文(Chinese), 翻譯(translation), 參考(reference), 尿素(urea), 代謝(metabolize), and 障礙(disorder) are in the Chinese monolingual dictionary, they would be grouped into words.

3.1.5 Extraction of Chinese translation candidate for the English OOV terms

The distance between the location of Chinese translation candidates and the English medical OOV term is important in this stage. This distance normally affects the possibility of a Chinese word to be a correct translation(Zhang, Huang et al. 2005). We first separate

the segmented web snippets into two parts: the part that comes before the English medical OOV term (front part) and the part that comes after the OOV term (back part).

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

1. We scan sequence of characters in the snippet to search for the first Chinese characters with in a predefined searching distance in front and back of the English OOV term because we discover that the Chinese translations usually occur not too far away from the English OOV term.

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

3. We extract all continuous Chinese characters after the nearest Chinese character. Detail of the process is shown in Fig. 3. 4.

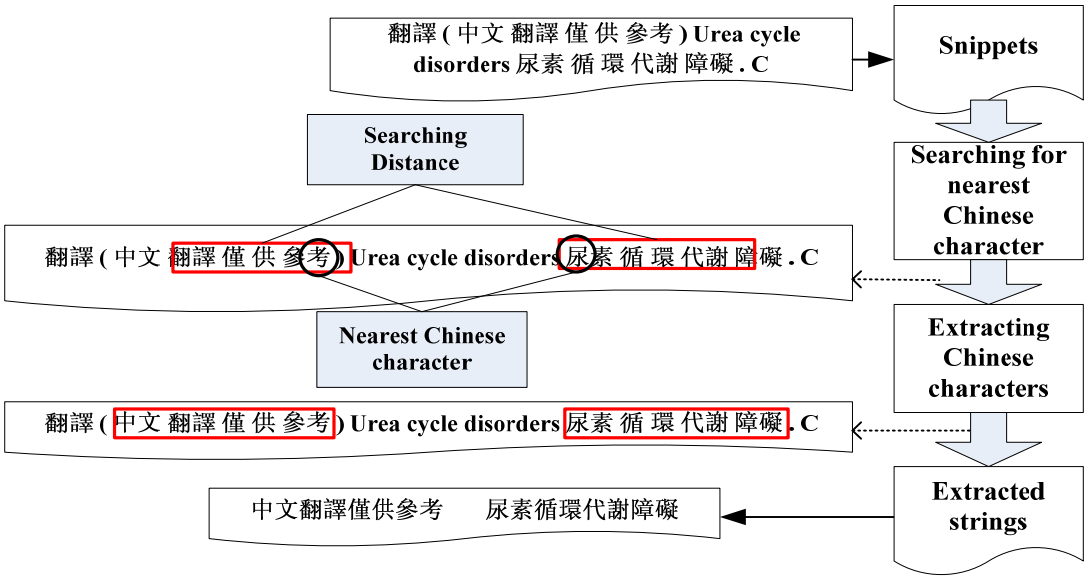

Figure 3.4 candidates extraction

In Fig. 3.4, “Urea cycle disorders” is an English medical OOV term; our system tries to find the nearest Chinese character in front and after the English OOV term. Then we extract continuous Chinese characters after the nearest Chinese character, they are: “中文翻譯僅供參考” and “尿素循環代謝障礙”.

3.1.6 Filter the Chinese translation candidate

There are three steps in this stage: first is scanning for possible candidates with medical keyword; second is generating possible candidates by using a technique of

longest common substring (LCS); and last is flittering possible candidates by using statistics and LCS.

Since we focus on English medical OOV terms only, there is a pattern for calling a Chinese disease. For the Chinese disease name, they usually end with certain medical keywords, such as "症" and "病" meaning “sickness”, "候群" meaning “syndrome”, "異常" and "缺陷" meaning “disorder.” We generate a group of Chinese medical term keywords from publicly available list of diseases (Health 2009).

By these keywords and longest common substring (LCS) algorithm (Lin 2004), we scan through snippets on both front part and back part to generate the co-occurrence frequencies for the possible translation candidates of the English medical OOV term. Different from the normal LCS where it selects the largest frequency and longest length, we keep all possible translation candidates ending with the keywords.

Since we kept all the possible translation candidates, there are a great number of candidates. We use average of the candidate’s occur frequencies to filter out some of the least possible candidates. For each English OOV term, the possible translation candidates after filter can be categorized into groups based on its substring component. For each of the group, only the one that has the longest length is selected according to the LCS algorithm.

3.1.7 Selecting the Chinese translation candidates

One way to select the Chinese translation candidates is to use association measures and co-occurrence frequencies. It is composed of four steps: selecting certain Chinese translation by co-occurrence frequencies; fetching the Chinese translations with its associated English OOV terms back to web; calculating association measures of the Chinese translations and English OOV terms; and selecting the correct Chinese translation by using association measures.

Our assumption is that the one Chinese translation cannot be the translation of two or more English medical OOV terms. For each Chinese translation candidates, we scan for it’s co-occur frequencies with each of its related English OOV terms. We select the pairs of Chinese translation candidate and English OOV term with the highest co-occurrence above a predefined co-occurrence score. We also remove any other pairs for the same Chinese translation candidate, showing in the table 3-1 below.

Table 3-1 An example of candidate selection based on co-occurrence frequencies

English OOV	Chinese translations	Co-occurrence frequencies
Urea cycle disorders	尿素循環代謝障礙	**30**
Urea cycle disorders	抗胰蛋白酶缺乏症	**8**
~~**Urea cycle disorders**~~	~~瓜 胺 酸 血症~~	~~**6**~~
Citrullinemia	瓜 胺 酸 血症	**26**
~~**Citrullinemia**~~	~~尿素循環代謝障礙~~	~~**15**~~

The association measures, which are popular and well-researched methods, can be used for discovering interesting relations between these OOV terms and their Chinese translation candidates (Hand, Mannila et al. 2001; Pei, Han et al. 2001; Viriyayudhakorn, Theeramunkong et al. 2008). Association measures are used to check the following two conditions:

1. When English medical OOV term exists in the webpage, what the relation for each Chinese translation candidate with the English medical OOV term is.

2. When each Chinese translation candidate exists in the webpage, what the relation for the associated English medical OOV term with the Chinese translation candidate is.

In this step, we query 1) English medical OOV term, 2) each Chinese translation candidate, 3) English medical OOV term and each Chinese translation candidate together to the web again to the Internet to collect the number of pages containing each of them. By getting the returned number of pages, we can build up a large set of information of how each OOV term and its translation candidates occur in the website. Note that we limited language of the snippets to Chinese only.

We employ two association rule measures (Hand, Mannila et al. 2001; Pei, Han et al. 2001) for our system; the Lift or Interestingness and the Conviction. Those association measures calculations are show as follows:

$$supp(c_1 \dots c_n) = \frac{S(c_1 \dots c_n)}{N} \tag{3.1}$$

$$supp(Eoov) = \frac{S(Eoov)}{N} \tag{3.2}$$

$$supp(Eoov \cup (c_1 \dots c_n)) = \frac{S(Eoov \cup (c_1 \dots c_n))}{N} \tag{3.3}$$

$$supp(c_1 \dots c_n) = \frac{S(c_1 \dots c_n)}{N} \tag{3.4}$$

$$lift(Eoov \rightarrow (c_1 \dots c_n)) = \frac{supp(Eoov \cup (c_1 \dots c_n))}{supp(c_1 \dots c_n) \times supp(Eoov)} \tag{3.5}$$

$$conv(Eoov \rightarrow (c_1 \dots c_n)) = \frac{1 - supp(c_1 \dots c_n)}{1 - \frac{supp(Eoov \cup (c_1 \dots c_n))}{supp(Eoov)}} \tag{3.6}$$

where $S(Eoov)$ is the number of pages containing OOV term, $S(c_1 \dots c_n)$ is the number of pages containing translation candidates that are composed of string $c_1 \dots c_n$, $Supp(Eoov \cup (c_1 \dots c_n))$ is the number of pages containing both OOV term and its translation candidate. *N* is the number of total web pages in the world.

Some Chinese translation candidates are selected in section 3.1.7, for any unselected Chinese translations, we pick up the Chinese translation with the highest Lift score and the lowest Conviction score. Both highest Lift score and lowest Conviction indicates that a Chinese translation occurs more often with the English OOV term.

Detailed experiment of this research is explained later on chapter 4.

3.2 Rules Based Candidate Generation with Data mining approach

We introduce this approach with a system design and detailed method procedures.

3.2.1 System design

The objective of this approach is to build an intelligent and automatic translation system to translate English medical OOV terms to Chinese medical terms. Since some Chinese translations of English medical OOV terms are not perfectly translated for example those contain English, symbols and numbers, we need a more intelligent approach to solve this problem. All existing English to Chinese translation approaches assumes Chinese translations have only Chinese characters, as a result English numbers or symbols, which sometimes are important in the Chinese translations, are ignored. This problem usually does not occur in name entity OOV terms but it becomes problematic for technical and medical type OOV terms. In this approach, we solve the Chinese translations of English medical OOV terms with a more intelligent approach. The extracting strings from the retrieved snippets do not only include the Chinese characters but also the English characters, numbers and symbols. Our rule based candidate generation approach is used to detect the boundary by specifying the type of characters that need to be included.

Quantifying the location distance between English OOV term and Chinese translation candidates is necessary because the distance can imply the possibility of being a correct translation. Association measures are proved to be useful in pervious approaches(Viriyayudhakorn, Theeramunkong et al. 2008; Qu, Viriyayudhakorn et al. 2009). Semantic conditional probability (SCP) (Silva, Ga et al. 1999; Cheng, Teng et al. 2004; Lu, Xu et al. 2007) could also help us to clarify which candidates are more likely to be a word. In order to integrate the distances, frequencies, SCP and association measures all together as features, we apply decision tree to select the correct translation

candidates.

According to our observation, some English OOV terms are not perfectly translated into Chinese, due to a special type of Chinese translation candidates which are the mixture of Chinese and non Chinese characters. At the present, none existing English to Chinese OOV term translation system can handle the translations. Some examples of Chinese translations of English medical OOV term is shown in Table 3-2

Table 3-2 Few examples of Chinese translation of English medical OOV term

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

3.2.2 Procedures

In this section, we discuss the automatic extraction of Chinese translation of English medical OOV terms. We apply eight steps to extract the correct Chinese translation of the English medical OOV terms, the steps are described as follows:

1) Extracting snippets containing English medical OOV terms from the web

2) Determining the boundaries of the web retrieved text

3) Extracting Chinese translation candidate for the English OOV terms

4) Extracting features of each Chinese translation candidates

5) Filtering the Chinese translation candidate

6) Applying the decision tree on each Chinese candidates

7) Selecting the Chinese translation candidates

A flowchart of this procedure is showed in Fig. 3.5.

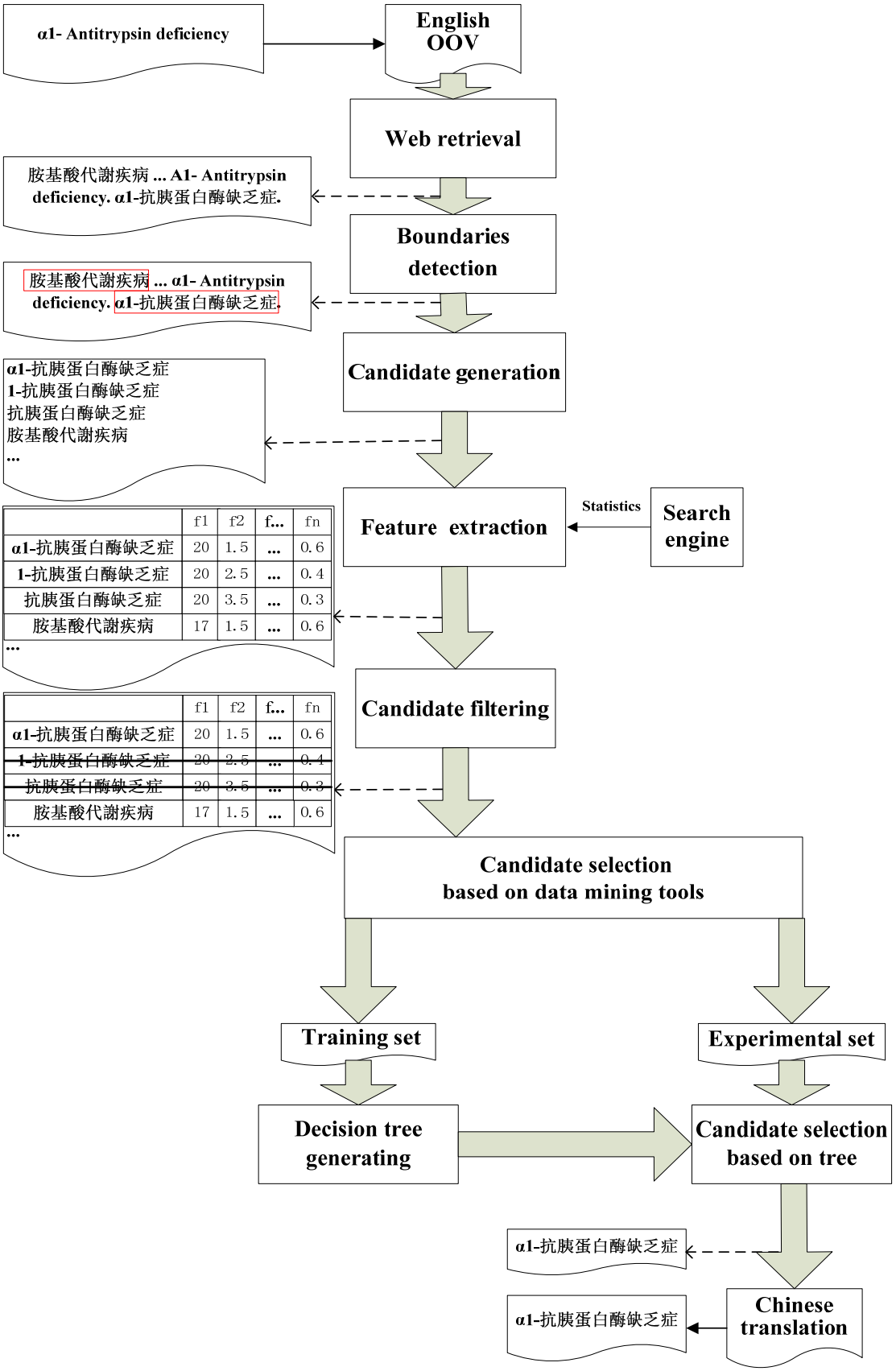

Figure 3.5 Flow chart of rule based candidate generation with data mining approach

3.2.3 Extracting English medical OOV terms from the web

Similar to the pre-processes like described in section 3.1.2 are performed. Unlike the previous approach, we separate the punctuations from any characters in the retrieved snippets. However, we do like to use a different example of snippets in order to show Chinese translation contains non-Chinese characters.

Urea cycle disorders & α1-Antitrypsin deficiency **(Title)**

Amino acid metabolicdisorders (Aminoacidopathies) 胺基酸代謝疾病. 260. 70. α1-Antitrypsin deficiency. ICD9. α1-抗胰蛋白酶缺乏症. 277.6. 118. Tyrosinehydroxylase deficiency. 酪胺酸羥化酶缺乏症 ... **(Summary)**

Http://www.nhicb.gov.tw/nhicbd00/ths68... **(URL)**

Figure 3.6 Example of the retrieved snippet for query "α1-antitrypsin deficiency"

3.2.4 Determining the boundaries of the web retrieved text

The distance between the Chinese translation candidates and the English medical OOV term is important in this approach because the closeness of the Chinese candidates implies the high possibility of the correct translation candidates(Cheng, Teng et al. 2004; Zhang, Huang et al. 2005). To find the distance, we separate the web snippets into two parts, the part that comes before the English medical OOV term and the part that comes after the OOV term.

The reason for all existing approaches choose not to include English characters, numbers and symbols for Chinese translation candidates is that those inclusions could add lots of incorrect translations, and the final result could be even worse than just extracting Chinese characters. Some examples of extracting only Chinese character compare to extracting all characters are shown below:

Since we would like to handle Chinese translation that contains English characters, numbers and symbols, we classify the characters in the web retrieved snippets into four different types. These types are English alphabets, Chinese characters, numbers, and symbols, which include punctuations. For each English OOV term, we check the type of characters exist in the English OOV term. Depending on the different types of characters in the English OOV term, we categorize the English OOV term into four different groups. They are as follows in Table 3-3:

Table 3-3 examples of extracting only Chinese character compare to extracting all characters

English OOVs	Correct Chinese translations	Snippets	Extracting Chinese characters only	Extracting all characters
Cystinosis	胱胺酸病	Urea Cycle ... Cystinosis 胱胺酸病 10.	胱胺酸病	胱胺酸病 10
phenylalanine hydroxylase	苯丙胺酸羥基化酶	苯丙胺酸羥基化酶 PAH: (phenylalanine hydroxylase)、和 BH4 輔脢的	苯丙胺酸羥基化酶	苯丙胺酸羥基化酶 PAH
GM1/GM2 gangliosidosis	GM1/GM2 神經節苷脂儲積症	GM1/GM2 神經節苷脂儲積症 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
α1- Antitrypsin deficiency	α1-抗胰蛋白酶缺乏症	270.6. 891207 ... icd9α1- Antitrypsin deficiency. α1-抗胰蛋白酶缺乏症.	抗胰蛋白酶缺乏症	Icd9α1-抗胰蛋白酶缺乏症
X-Linked Hypophosphatemic Rickets	性聯遺傳低磷酸佝僂症	而 ... X-Linked Hypophosphatemic Rickets;XLH 性聯遺傳低磷酸佝僂症.	性聯遺傳低磷酸佝僂症	XLH 性聯遺傳低磷酸佝僂症
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	755.59. 45. DiGeorge's syndrome. DiGeorge's 症候群.	症候群	DiGeorge's 症候群.
	8/8 correct		**5/8 correct**	**2/8 correct**

1. Group 1: the English OOV term contains only English alphabets;
2. Group 2: the English OOV term contains English alphabets and symbols;
3. Group 3: the English OOV term contains English alphabets and numbers;
4. Group 4: the English OOV term contains English alphabets, numbers, and symbols.

For each of the above groups, we can choose the types of characters we prefer. For simplicity, these preferred types of character are called *potential key characters*. Since we have four different groups of English OOV term, we can apply them into four different conditions of potential key characters. These conditions as follows:

1. If the English OOV term belongs to group 1, then the potential key characters contain English alphabets and any Chinese characters;
2. If the English OOV term belongs to group 2, then the potential key characters contain English alphabets, any Chinese characters and the same symbols;

3. If the English OOV term belongs to group 3, then the potential key characters contain English alphabets, any Chinese characters and same numbers;

4. If the English OOV term belongs to group 4, then the potential key characters contain English alphabets, any Chinese characters, same numbers and the same symbols. Details of the rules are shown in Table 3-4.

Table 3-4 Rules for Potential key characters

Type of characters in English OOV term	Potential key characters
English string of characters	● English string of characters ● Chinese string of characters
English string of characters and Numbers	● English string of characters ● Chinese string of characters ● Same numbers
English string of characters and Symbols	● English string of characters ● Chinese string of characters ● Same symbols
English string of characters, Symbols, and Numbers	● English string of characters ● Chinese string of characters ● Same symbols ● Same numbers

By using the above conditions, we detect the boundaries of the snippet in front and behind the English OOV term according to the following steps.

1. We scan for each characters and words to check for the first Chinese characters within a predefined searching distance because we discover that the Chinese translations usually do not occur too far away from the English OOV term.

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

3. We extract all potential key characters from both sides of the nearest Chinese character. Details are shown in Fig. 3.7.

3.2.5 Extracting Chinese translation candidates for the English OOV terms

Since the section 3.2.4 only identifies the boundary of the translation candidates, all possible patterns of outputs in both the front string and back string are considered in this step, for example a string with five characters/words “Robinow 氏症候群”, it generates

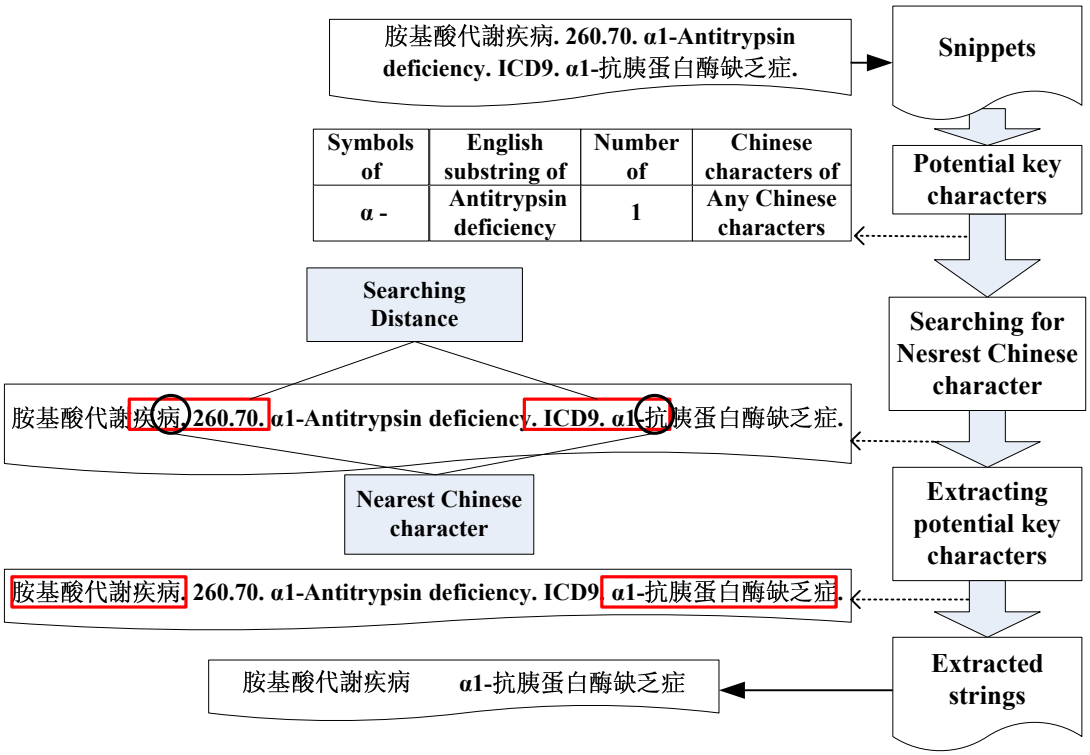

Figure 3.7 boundaries detection

An example of boundary detection is shown in Fig. 3.7. In this example, the English OOV term is “α1- Antitrypsin deficiency”, we found the nearest Chinese character in front of the English OOV term is “病”, and the nearest Chinese character behind the English OOV term is “抗”. Then, we extract for any continuous potential key characters from both sides of these initial Chinese characters. The extractions stopped at the full stop “.” because it is not a potential key character.

substrings as Robinow,氏,症, 候, 群, Robinow 氏, Robinow 氏症, Robinow 氏症候, 氏症, 氏症候, 氏症候群, 症候, 症候群, 候群 and Robinow 氏症候群. Out of thesesubstrings, we focus on English medical OOV terms only, for the translation in Chinese, they usually end with certain keywords (Chen, Yang et al. 2003), such as sickness "症" or "病", syndrome "候群", disorder "異常" or "缺陷". We constructed a list of keywords for Chinese medical terms from publicly available list of diseases(Health 2009). We generate the possible Chinese candidates ended with certain medical keywords.

3.2.6 Extracting features of each Chinese translation candidate

In this section, we describe the details of each feature we extract from each Chinese translation candidate.

3.2.6.1 Frequency

The frequency represents a very important statistical feature for the retrieved Chinese translation candidates. The more frequent is a Chinese translation co-occurs with an English OOV term, the more likely it is the correct translation. We collect the frequency of the Chinese translation candidates by simply counting the number of times those translation candidates occur in our retrieved snippets. There are three different types of frequencies of each Chinese translation candidates depending on the locations that are found, the *front frequency* and the *back frequency*. In addition, we also keep track the total frequency for each Chinese translation candidates.

3.2.6.2 Front and back distance of Chinese translation candidates

When a Chinese translation candidate appears close to its inputting English OOV term, it is more likely that Chinese translation is correct (Cheng, Teng et al. 2004; Zhang, Huang et al. 2005). The distance between the locations of English OOV term and each Chinese translation candidate can be calculated by counting the number of characters that Chinese translation candidate is away from the inputting English OOV term in the web retrieved snippets. For each candidate, it appears in more than one snippet, and the distance for this candidate in each snippet is different. We use the average distance of each candidate. Details are shown in Fig. 3.8.

2009 年 1 月 10 日 ... (α1-antitrypsin deficiency) α1-抗胰蛋白酶缺乏症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)
(Aminoacidopathies) 胺基酸代謝疾病. 260. 70. α1-Antitrypsin deficiency. ICD9. α1-抗胰蛋白酶缺乏症. 277.6. 118. Tyrosinehydroxylase deficiency. 酪胺酸羥化酶缺乏症 ...(Summary)

Candidates	Average distance
α1-抗胰蛋白酶缺乏症	4.5
1-抗胰蛋白酶缺乏症	5.5
-胰蛋白酶缺乏症	6.5

Figure 3.8 Example of distance

An example of average distance calculation is shown in Fig. 3.8, where “2009 年 1 月 10 日 ... (α1-antitrypsin deficiency) α1-抗胰蛋白酶缺乏症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)" and "(Aminoacidopathies) 胺基酸代謝疾病. 260. 70. α1-Antitrypsin deficiency. ICD9. α1-抗胰蛋白酶缺乏症. 277.6. 118. Tyrosinehydroxylase deficiency. 酪胺酸羥化酶缺乏症 ...(Summary)" are snippets. The distance between "α1-antitrypsin deficiency" and "α1-抗胰蛋白酶缺乏症" in the first snippet and second snippet are 1 characters and 8 characters, respectively. The average distance for this candidate is 4.5 characters.

3.2.6.3 Symmetrical Conditional Probability

We calculate the SCP score for each possible Chinese translation candidate. Symmetrical Conditional Probability (SCP) checks each character and substring in the possible Chinese translation. By calculating the frequency of each substring in the corpus and comparing them to the frequency of the Chinese translation. SCP score is higher when the substring of the Chinese translation occur less often in the corpus than it occurs only within the translation itself. If a translation has a high SCP score, the translation is more likely to be a word not a sentence.

3.2.6.4 Association Measures and chi-square

Association measures are proved to be useful in previous work (Viriyayudhakorn, Theeramunkong et al. 2008; Qu, Viriyayudhakorn et al. 2009), so applying them to find the relationship of each Chinese translation candidate to its associated English OOV term would be reasonable. The association measures between English OOV term and its Chinese translation candidates can be found by counting the number of pages returned by querying English OOV term or Chinese translation candidates or English OOV term together with Chinese translation candidates to the web. The conventional association measures explained in Chapter 2 request user to define the number of total data, which is the existing total number of web pages. We came up with a variant of association measures that does not use the total number of existing web pages. Those association measure formulas are explained below.

$$Supp(Eoov \rightarrow (c_1 \dots c_n)) = S(Eoov \wedge (c_1 \dots c_n)) \tag{3.7}$$

$$Conf(Eoov \rightarrow (c_1 \dots c_n)) = \frac{S(Eoov \wedge (c_1 \dots c_n))}{S(Eoov)} \tag{3.8}$$

$$lift(Eoov \rightarrow (c_1 \dots c_n)) = \frac{S(Eoov \wedge (c_1 \dots c_n))}{S(Eoov)S(c_1 \dots c_n)} \tag{3.9}$$

$$Conv(Eoov \rightarrow (c_1 \dots c_n)) = \frac{S(Eoov)(S\neg(c_1 \dots c_n))}{S(Eoov \wedge \neg(c_1 \dots c_n))} \quad (3.10)$$

where $Eoov$ is the English OOV term, $(c_1 \dots c_n)$ is the generated Chinese translations, and S is a function that takes a query as the input and returns the number of pages retrieved from the web. Note that conviction cannot be calculated due to missing $(S\neg(c_1 \dots c_n))$, we explain how we solve that below.

Since search engines do not offer us the number of pages not including$(c_1 \dots c_n)$, so it can only be approximated. Since our translation candidates are OOV terms, they occur infrequently in web pages compared to the total number of web pages in the Internet, so the number of web pages that does not have any $(c_1 \dots c_n)$ is assume to be equal to the total number of web pages which equals to one. The association rule for conviction is modified again as follow.

$$Conv(Eoov \rightarrow (c_1 \dots c_n)) = \frac{S(Eoov)}{S(Eoov \wedge \neg(c_1 \dots c_n))} \quad (3.11)$$

We get all four association measure scores for each set of Chinese translation candidates with its English OOV term. We then calculate the chi-square for the English OOV term and its Chinese translation candidates.

3.2.7 Filtering the Chinese translation candidates

Since there are many Chinese translation candidates to be processed, some statistical ways to filter the Chinese translation candidates are performed. In each set of Chinese translation candidates from an English OOV term, we choose the related Chinese translation candidates that are in the top 70% from the frequency rank, the threshold of 30% is selected because our previous work (Qu, Viriyayudhakorn et al. 2009) suggests that threshold has the best recall rate.

However, this would still offer us more than 100 Chinese translation candidates for each English OOV term. We can reduce this number by simply taking the shortest distance between English OOV and translation candidates.

3.2.8 Applying the decision tree on each Chinese candidates

Data mining approaches have been proven to be more efficient than rule-based

approach in CLIR (Limungkura, Theeramunkong et al. 2009), number of research suggest that data mining outperforms rule-based system in multiple ways. We apply decision tree as our machine learning system because decision tree is one of the most powerful and popular tools for classification and prediction. Since decision tree is a supervised learner, we need to label a portion of data to create a training data set. We manually compare the results from our generated Chinese translation candidates with the list of translations provided by Taiwanese government to create a training data set. Then, we apply the decision tree to the labeled training data set to generate the model tree. Since our features including association measure, chi-square, front and back distance, frequency, and SCP, it is difficult for human to decide which feature is more important than others, so we apply a brute force feature selection method to decide which feature is more important than others by testing all the combination of features. We explain how to apply the generated model tree to predicate the correct Chinese translation in next section.

3.2.9 Selecting the Chinese translation candidates

Selection process or experimental process is straight forward based on the generated tree we run the testing set of the translation candidates on the model tree and pick up the ones that are correct according to the model tree. However, there are possibilities that one English OOV term can match to more than one Chinese translation candidates as the correct translation candidates according to the decision tree. Among those selected Chinese translation candidates, some maybe incorrect translations. In order to filter them out, we use the prediction confidence of each Chinese translation candidates to select the unique correct Chinese translation candidates.

Detail experiments are discussed in Chapter 4.

Chapter 4

Experiments and results

In this chapter, we describe the experiments and results for two different approaches. They are LCS candidate generation with association measures approach and rule based candidate generation with data mining approach. This chapter is organized as follows. We described how experiments are setup and results for the LCS candidate generation with association measures approach in Section 4.1 and for the rule based candidate generation with data mining approach in Section 4.2.

4.1 Using LCS Candidate Generation with Association Measures.

In this section, we describe the experiments and results in details for the LCS candidate generation with association measures approach presented in section 3.1. This section is organized as follows: we describe the experiment setup, evaluation scheme, and experimental results, in 4.1.1, 4.1.2, and 4.1.3, respectively.

4.1.1 Experiment setup

In this section, we describe how the data are prepared, translation candidate generation, and translation candidate selection. The details are below.

4.1.1.1 Data preparation

A list of English special medical OOV terms from the Taiwanese centers for disease control, R.O.C. (Taiwan 2009) was used for our experiments. The list is called “List of announced rare disease of ICD-9-CM” (ICD9 2009). It contains a total of 184 ICD9 special medical terms, which are presented in both English and Chinese. For each of those English OOV terms, we query them to the search engine to retrieve the snippets. We chose Yahoo API for our search engine because Yahoo API (Yahoo 2009) accepted automatic queries from a computer where most other search engines blocked our IP. After querying those English OOV term to the Yahoo API, we were able to retrieve a total of 7,456 snippets from 1,133 different domains. A table showing top 50 domains is shown in appendix A. For each retrieved snippets we segment it by using DEDE Chinese segmentation system (DeDe 2009) due to its free availability and extensibility. The monolingual dictionary provided by this segmentation tool contains 112,492 Chinese words. A list of medical keywords for Chinese special medical term was extracted from

the Taiwanese Government Web Services for Genetic Diseases (Health 2009). This list of medical keywords is used later for translation candidate generation.

4.1.1.2 Translation candidate generation process

For each segmented snippets obtained from the previous section, we divide them into two parts, the part in front of the English OOV term and the part in the back of the English OOV term. For both parts, we scan each character one-by-one within the limited distance to find the closest Chinese character to the English OOV term. For simplicity, we call this distance the *searching distance*. In our experiment, the searching distance is set to be 8 characters, because we discovered any distance equal or more than 8 would result the best recall rate, Fig. 4.1 shows the recalls of different searching distance. If no Chinese character is found, the search stops. Otherwise, we extract all contiguous Chinese characters. For those extracted Chinese characters, we use a list of medical keywords mentioned in the previous section to generate translation candidates, because Chinese medical names usually end with certain medical key words.

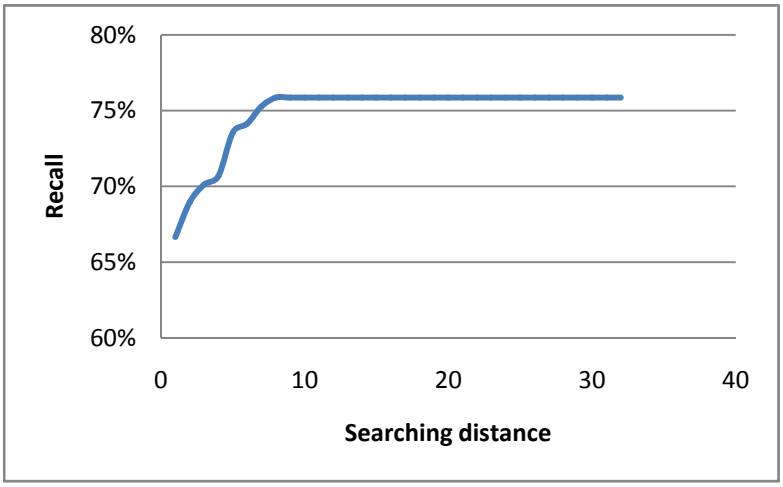

Figure 4.1 Recalls of different searching distance for LCS generated candidates

4.1.1.3 Candidate selection process

From the translation candidates generated from the previous section, we use the average of candidate's frequencies to remove candidates that are not likely to be the translations. Note that we exclude all translation candidates with frequencies of 1, since Lu suggests that translation candidates with very low frequencies such as 1 are not the correct translation candidates (Lu, Xu et al. 2007).

After filtering for each possible Chinese translation candidates we firstly select distinct pairs of Chinese translation candidate and English OOV term with the highest co-occurrence above the co-occurrence score of 10, because Zhang suggest that candidates with high co-occurrence frequencies such as 10 are likely to be a correct

translation(Zhang and Vines 2004). We secondly use association measures to select the most appropriate Chinese translation candidates. We get numbers of returned pages by querying the English OOV term, Chinese translation candidates and both of them together to the Internet. We use those numbers of returned pages to calculate two association measures: Lift and Conviction. For Lift, we select the Chinese translation candidates with the highest Lift score, because the higher Lift score the higher the co-occur relationship between the English OOV term and the Chinese translation. For Conviction, we select the Chinese translation candidates with the lowest Conviction score, because the higher the conviction score the lower the co-occur relationship between the English OOV term and the Chinese translation.

4.1.2 Evaluation scheme

The ultimate goal is to get the Chinese translation of the English medical OOV term, but the process involves three different steps, they are: web retrieval of English OOV term; Chinese translation candidate generation; and translation candidate selection. We would have to evaluate these steps separately, and then present the overall performances. A way of define the correct Chinese translation candidates is explained below.

From the data preparation process, we have the correct Chinese translation for each input ICD9 English special medical OOV terms provided by the Taiwanese government. We can evaluate the correctness of our results by comparing the results with those correct Chinese translations. If our Chinese translation is semantically same from the data preparation, we call it a “correct match”. If our translation is semantically different from the data preparation, we call it a “wrong match”.

In order to evaluate our approach, we compare the results from our approach to the length co-occurrence approach invented by Zhang (Zhang and Vines 2004).We use precision and recall to compare our results with length co-occurrence results. The calculation of precision and recall are shown as follows.

$$Precision = \frac{relevant\ documents\ \cap\ retrieved\ documents}{retrieved\ documents} \tag{4.1}$$

$$Recall = \frac{relevant\ documents\ \cap\ retrieved\ documents}{relevant\ documents} \tag{4.2}$$

4.1.3 Experimental results

Since some English medical OOV terms do not have the Chinese translation in web snippets, their translations were lost during the web retrieval process. For our experiment, although the input is 184 English OOV terms, the web retrievable OOV terms are only

179 terms. Among that, 177 do have the correct Chinese translation in the retrieved snippets when they were checked manually. Therefore, we use 177 as baseline for total recall. Our LCS based candidate generating system generated a large set of Chinese translation candidates. We can find the correct translation for 132 English OOV terms among those candidates if we manually check these candidates. The length and co-occurrence method also generated a large set of Chinese translation candidates; we can find the 118 correct Chinese translations if checked manually. The recall for candidate generation of existing approach and our approach are shown in Fig. 4.2.

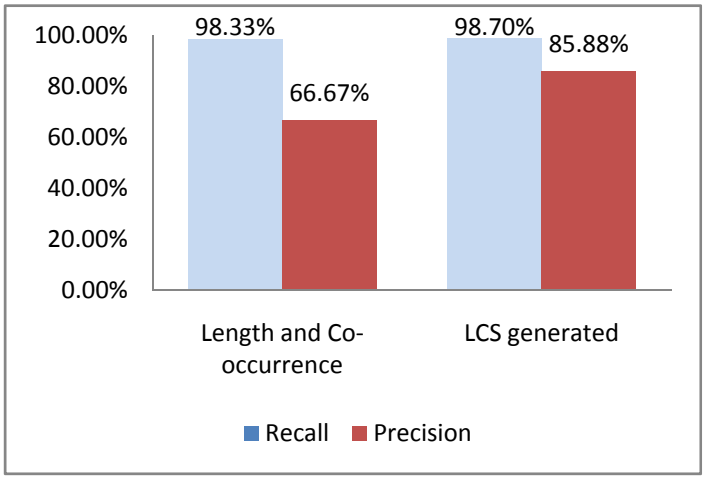

Figure 4.2 shows the recall and precision of Length and Co-occurrence candidate generation compare to LCS candidate generation

For the candidate selection part, the length co-occurrence approach, the overall precision is 64.41%, our overall precision increase significantly to 84.18%. Results of experiment are shown in Fig. 4.3.

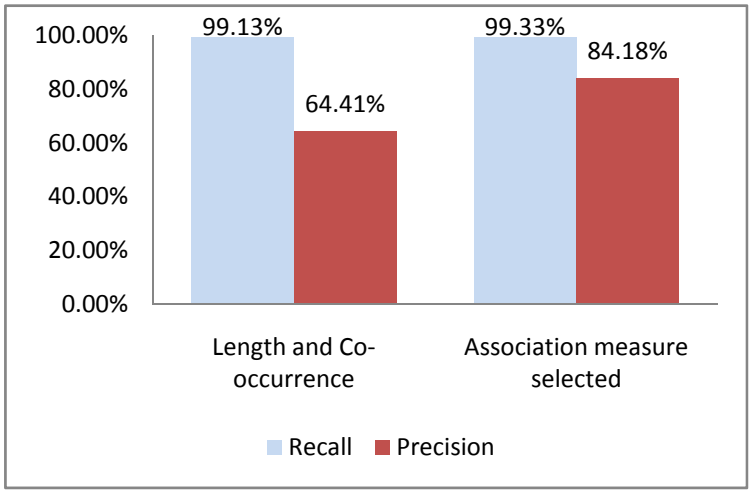

Figure 4.3 shows the overall recall and precision of Length and Co-occurrence compare to our proposed method

We compare our results using two different association measures with length co-occurrence approach in more detail. The lift and conviction give a very similar result for selecting candidate pairs. The lift out-performs all the rest methods; it reaches 84.18% correct matched terms. Among those results, the length co-occurrence approach has the lowest score when selecting correct matches. Results of this experiment is shown in Fig. 4.4

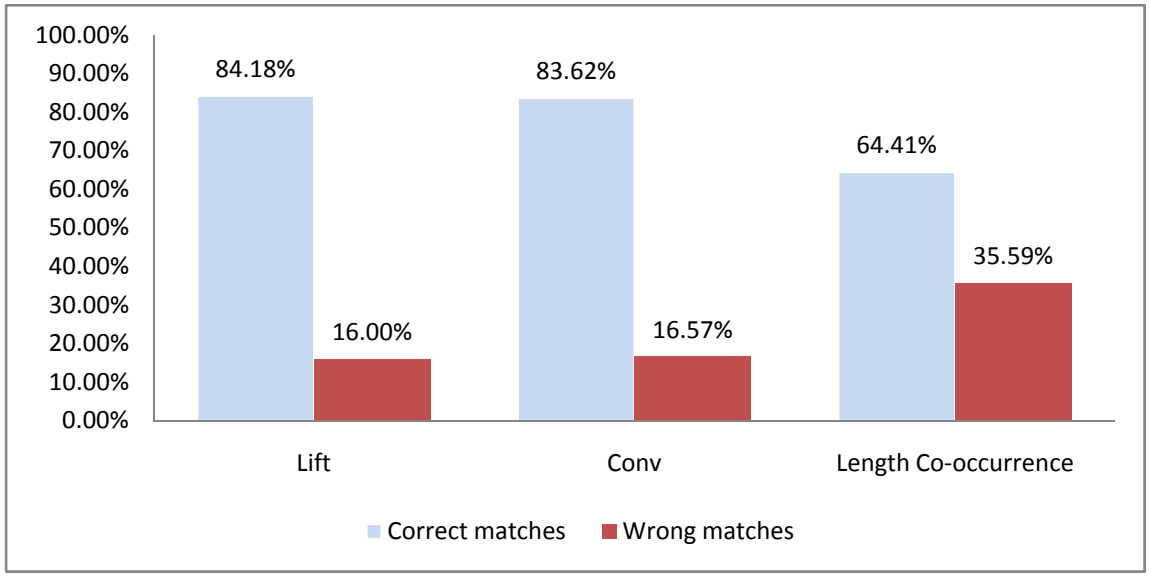

Figure 4.4 Comparison of each OOV term translation approach in great detail

4.2 Rule Based Candidate Generation with Data Mining Approach

In this section, we describe the detail experiments and results for the rule based candidate generation system and data mining approach, presented in section 3.2. This section is organized as follows: we describe the experiment setup, evaluation scheme, and experimental results, in 4.2.1, 4.2.2, and 4.2.3, respectively.

4.2.1 Experiment setup

In this section, we describe how the experiment is setup. This section is organized as follows: we describe the data preparation and web retrieval, translation candidate generation, and translation candidate selection, in 4.2.1.1, 4.2.1.2, and 4.2.1.3, respectively.

4.2.1.1 Data preparation

Similar to the previous experiment, the list of English OOV terms, the search engine, and the Chinese medical keyword list are all same with section 4.1.1.1.

4.2.1.2 Translation candidate generation process

For each English OOV term, we first check the contents of this English OOV term to define its potential key characters, presented in section 3.2.4. For each snippets obtained from the previous section, we divided them into two parts: the part in the front of the English OOV term and the part in the back of the English OOV term. We scan for each characters and words to check for the nearest Chinese characters within a *searching distance* of 8 characters because we discover that *searching distance* equal or above 8 characters performs the best recall, Fig. 4.5 shows the recalls using different searching distance. If any Chinese character is found within the searching distance, we call that Chinese character the nearest Chinese character and extract all continuous potential key characters on both side of this nearest Chinese character. Similar to the previous experiment, for those extracted potential key characters, we used a list of medical keywords from the previous section to generate translation candidates.

Figure 4.5 Recalls using different distance for rule based system generated candidates

4.2.1.3 Candidate selection process

We use some simple statistics to filter the possible translation candidates obtained from the previous section. We take the top 70% of the frequency rank and select the Chinese translation candidates, which are 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. Fig. 4.6 shows recalls and noises using different frequency rank.

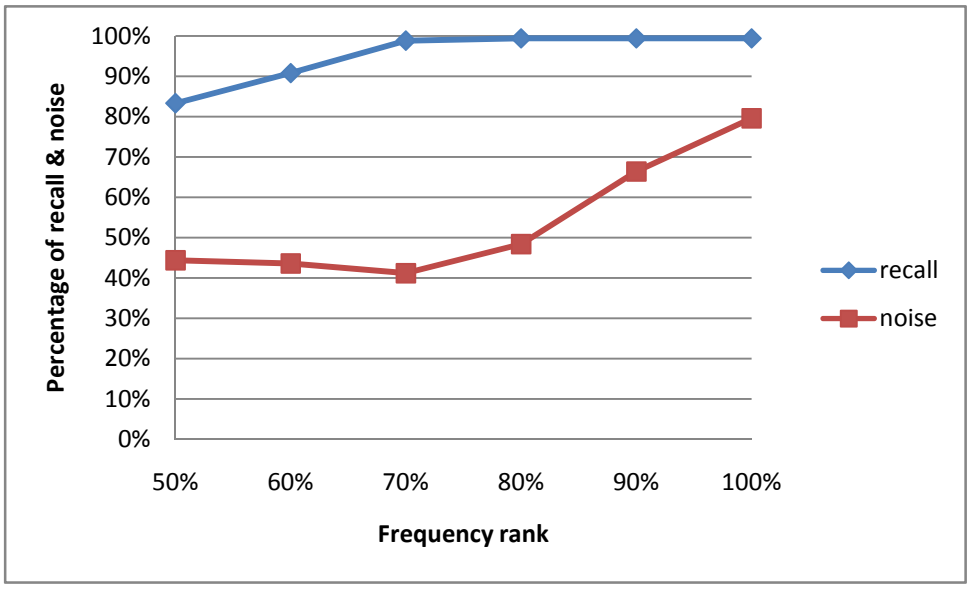

Figure 4.6 Recalls and noises using different frequency rank

From the filtered possible translation candidates we take 90% as the train set and 10% as the experimental set for the decision tree process, and we loop the trains set and experimental set through the entire data set.

A few features were extracted for each possible translation candidate after filtering. Those features are: 1) the front and back distances, 2)the co-occurrence frequencies, 3) four association measures (lift, conviction, confidence, and support) scores between each pair of English OOV term and its Chinese translation candidates, 4) chi-square scores between each pair of English OOV term and its Chinese translation candidates, and 5) SCP scores for each Chinese translation candidates. To select the correct translation candidates, we applied the decision tree to those candidates with features to generate the tree and we use the tree to select the correct translation. We choose Rapid Miner (Rapidminer 2009) to be the tool for data mining and error analysis because it offers error analysis for each individual pair of English OOV term and its Chinese translation candidates. We try to optimize the parameters used in the decision tree by adjusting each parameter. We find the best parameters for this experiment and they are shown in Table 4-1.

4.2.2 Evaluation scheme

Similar to the previous experiment, this experiment is divided into three different parts. They are the web retrieval part, the candidate generation part, and the candidate selection part. The evaluation scheme for the web retrieval part and the candidate generation part are same with section 4.1.2. For the candidate selection part, we check the number of translation candidates that are correct and were selected by decision tree approach. Since

Table 4-1 the decision tree parameters setting

Decision Tree Parameters	**Setting**
Criterion	Gain ratio
Minimal size for split	4
Minimal leaf size	2
Maximal depth	20
Confidence	0.25

Where Criterion specifies the used criterion for selecting attributes and numerical splits, Minimal size for split is the minimal size of a node in order to allow a split, Minimal leaf size is the minimal size of all leaves, Maximal depth is the maximum tree depth, and Confidence is the confidence level used for the pessimistic error calculation of pruning.

there are different features affecting the decision tree in the candidate selection process, we separate these features into two different groups: 1) the basic features like frequencies and distances and 2) the advanced features like SCP, association measures, and chi-square. For advanced features, we checked all possible combinations of those features and evaluated the combinations that yield high precisions. The method that we used for evaluation the correct Chinese translation candidates is same as section 4.1.2.

In order to compare our approach over existing approaches, we choose the length co-occurrence approach to experiment on the same data with our approach. We use precision, accuracy and recall to compare our results with length co-occurrence results. The calculation of precision and recall are same as section 4.1.2. In this experiment, we need some classification context to evaluate the decision tree performs. Calculations for accuracy, F-measure, precision, and recall for classification context are shown below:

$$Precision\ for\ classification = \frac{True\ positive}{True\ positive + False\ positive} \quad (4.3)$$

$$Recall\ for\ classification = \frac{True\ positive}{True\ positive + False\ negetive} \quad (4.4)$$

$$Accuracy = \frac{True\ positive + True\ negetive}{True\ positive + False\ positive + True\ negetive + False\ negetive} \quad (4.5)$$

$$Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4.6)$$

4.2.3 Experimental results

For our experiment, the web snippets are same with the section 4.1.3. Our rule based candidate generating system generated a large set of Chinese translation candidates. We can find the correct translation for 172 English OOV term among those candidates if we manually check these candidates. We found that this rule based candidate generation system could generate translation candidates at a high precision of 98.85%. The recall and precision for candidate generation of existing approach and our approach are shown in Fig. 4.7.

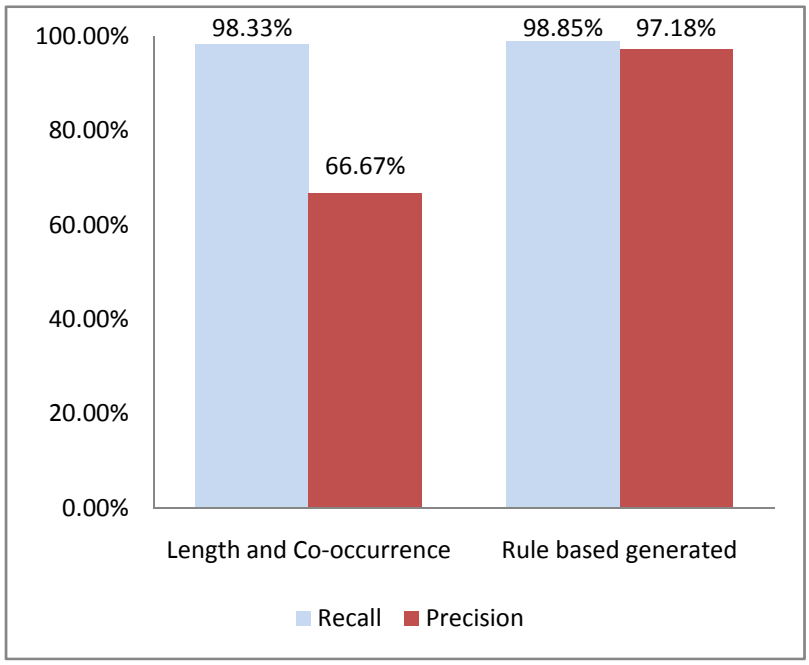

Figure 4.7 shows the recall and precision of Length and Co-occurrence candidate generation compare to rule based candidate generation

For the candidate selection process, there are a total of 275 pairs of Chinese translation candidates and its English OOV terms when applying to the decision tree. In order to test the advanced features, we compared the precisions of all combinations of advanced features, where basic features are always included in each combinations. Table 4-2 shows the expirment results for all combinations of advanced features.

In Table 4-2, TP, FP, FN and TN represents true positive, false positive, false negetive and true negetive in a standard confusion matrix, respectively. A, B, C, D, E, F represents SCP, lift, conviction, confidence, support, and chi-square, respactively.

Table 4-2 experiment results for candidate selection

Combinations	TP	FP	FN	TN	precision	recall	accuracy	F-measure
ABC, ABCD, ABCDE, ABCE	195	1	3	76	0.99	0.98	0.99	0.99
AB, BC, BCD, BCDE, ABD, ABDE, ABE, BCE	195	2	3	75	0.99	0.98	0.98	0.99
B, BD, BDE, BE	195	3	3	74	0.98	0.98	0.98	0.98
ACDE, CDE	194	6	4	71	0.97	0.98	0.96	0.97
all features, BCDEF, BDEF, BEF, ABCDF, ABCEF, ABCF, ABDEF, ABDF, ABEF, ABF, BCDF, BCEF, BCF, BDF	191	4	7	73	0.98	0.96	0.96	0.97
ADEF, AEF, AF, BF, DEF, DF, EF, F, ADF,	191	5	7	72	0.97	0.96	0.96	0.97
ACDEF, CDEF, CEF, CF, ACDF, ACEF, ACF, CDF	190	5	8	72	0.97	0.96	0.95	0.97
ACD, CD	194	14	4	63	0.93	0.98	0.93	0.96
ADE	194	18	4	59	0.92	0.98	0.92	0.95
CE	188	13	10	64	0.94	0.95	0.92	0.94
DE	194	19	4	58	0.91	0.98	0.92	0.94
ACE	184	12	14	65	0.94	0.93	0.91	0.93
AE	185	16	13	61	0.92	0.93	0.89	0.93
E	185	17	13	60	0.92	0.93	0.89	0.92
C	188	21	10	56	0.90	0.95	0.89	0.92
AC	185	20	13	57	0.90	0.93	0.88	0.92
A, AD, D	198	77	0	0	0.72	1.00	0.72	0.84

Table 4-2 is ranked according to accuracies. In the top part of the table where the high accuracies are located, we found out those combinations require lift as the feature. In the bottom part of the table where the low accuracies are located, we found out those combinations are resulted by SCP and confidence. We show few high accuracy tree graphs using association measures and chi-square as root in the decision tree as follows.

Figure 4.8 shows the best result when using Chi-square as root. The tree is up to 5 comparison nodes, features used in decision tree includes Chi-square, SCP, and frequency. The precision is up to 97.94%.

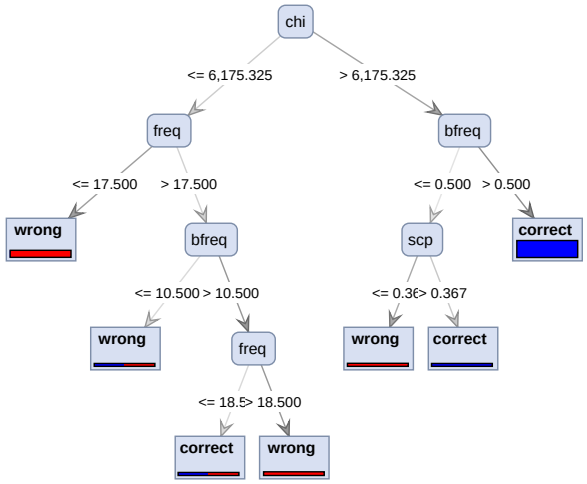

Figure 4.8 best result by Chi-square

where chi, freq, bfreq, and scp represents chi-square, total frequency, back frequency, and SCP scores, respectively. According to the above graph, most Chinese translation candidates with a chi-square score over 6175 and back frequency more than 0.5 are correct translations.

Table 4-3 confusion matrix by chi-square

Correct class	Wrong class	Classes / Classification
191	4	Classified as Correct
7	73	Classified as Wrong

Figure 4.9 shows the best result when using Conviction as root. The generated tree is quite complicated, which is up to 6 comparison nodes, features used in the decision tree include Conviction, Confidence, Support, SCP, and frequency. The precision is up to 97%.

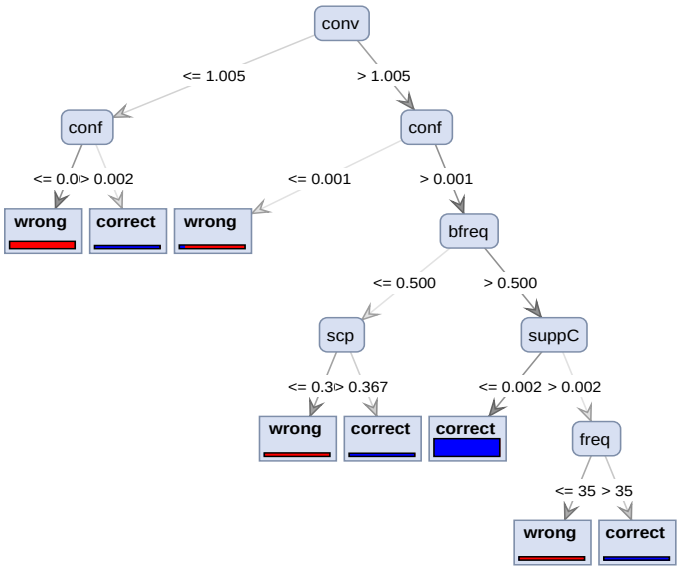

Figure 4.9 best result by Conviction

where conv, conf, freq, bfreq, scp, and suppC represents conviction, confidence, total frequency, back frequency, SCP scores, and support, respectively. According to the above graph, if conviction is selected as root by the decision tree, the majority of the correct Chinese translation candidates can be found after 4 comparison nodes.

Table 4-4 confusion matrix by Conviction

Correct class	Wrong class	Classes / Classification
194	6	Classified as Correct
4	71	Classified as Wrong

Figure 4.10 shows the best result when using Support as root. The generated tree is quite complicated which is up to 8 comparison nodes, features used in the decision tree include Support, Confidence, SCP, and frequency. The precision is up to 91.50%.

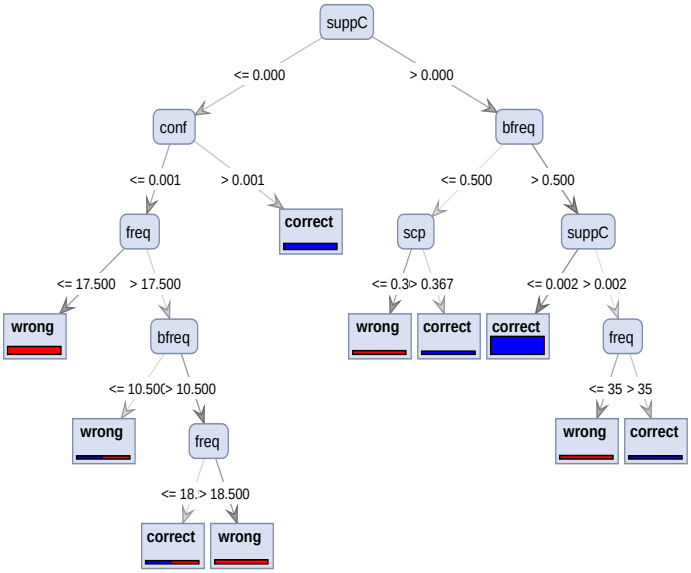

Figure 4.10 best result by Support

where suppC, conf, freq, bfreq, and scp represents support, confidence, total frequency, back frequency, and SCP scores, respectively. Although the tree graph is more difficult than the one using conviction as root, but when using support as root, decision tree only took 3 comparison nodes to classify the majority of the correct Chinese translation candidates.

Table 4-5 confusion matrix by Support

Correct class	Wrong class	Classes / Classification
194	18	Classified as Correct
4	59	Classified as Wrong

Figure 4.11 shows best result when using Lift as root. This tree is also the overall best result for the candidate selections process. The generated tree has only 3 comparison nodes; features used in the decision tree include Lift, SCP, and frequency. The precision is up to 99.48%.

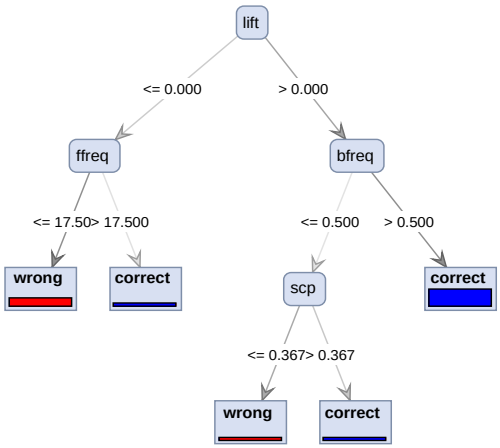

Figure 4.11 over all best decision tree resulted by Lift

where lift, ffreq, bfreq, and scp represents lift, front frequency, total frequency, back frequency, and SCP scores, respectively. When using lift as root, decision tree took only 2 comparison nodes to classify the majority of the correct Chinese translation candidates.

Table 4-6 confusion matrix by Lift

Correct class	Wrong class	Classes / Classification
195	1	Classified as Correct
3	76	Classified as Wrong

In order to check our rule based candidate generation and data mining approach over existing approaches, we compute the best decision tree result into our candidate generation system and check how many candidates are correctly translated. By compare it with length Co-occurrence result. Out of 177 English OOV terms, our approach can translate 169 terms. We have a much higher overall OOV term translation precision of 95.48%, shows in Fig. 4.12.

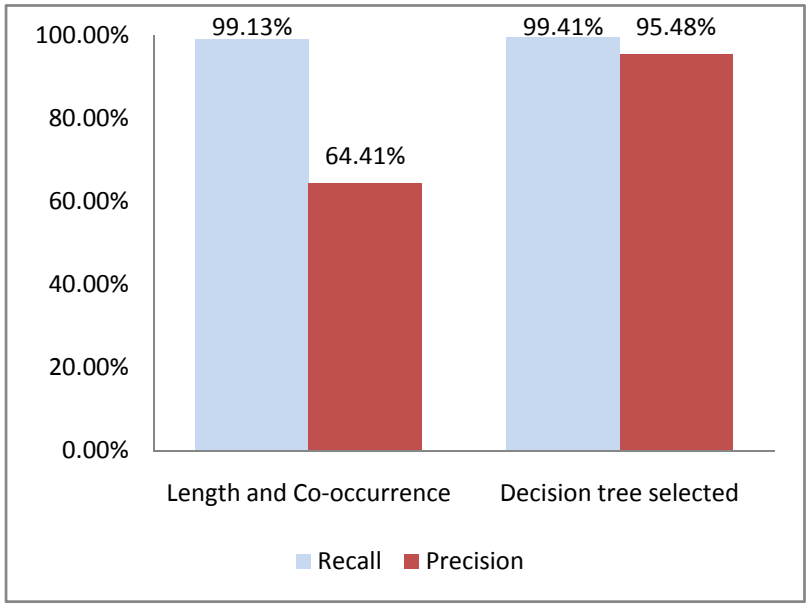

Figure 4.12 Comparison of overall OOV term translation recall and precision of Length and Co-occurrence to our rule based candidate generation and machine learning approach

Chapter 5

Discussions

In this chapter we provide explanations and discussions of behaviors of results obtained from our two approaches presented in the previous chapter. 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: Section 5.1, we discuss the results from experiments and also explanation for longest common substring candidate generation with association measure approach. Section 5.2, we do the same thing for rule based candidate generation with data mining approach. Section 5.3, we compare the above two approaches.

5.1 Discussion for LCS candidate generation with association measure approach

The association measure approach has a major advantage over existing methods because it selects the translation candidate automatically based on their association measures scores. Most existing method requires human to select the correct translation candidates.

In the experiment of association measure approach, the precision and recall reach up to 84.18% and 85.88%, respectively. Some long Chinese translated candidates that contain other unrelated Chinese words in front of the correct translation candidates cause the incorrect translations. When we fed these long and incorrect Chinese translation candidates back to the web, the association measure becomes zero. The incorrect Chinese translated candidates usually become a problematic in the LCS candidate generation system. Another source of error is from punctuations and other non-Chinese characters that may be embedded as part of English ICD9 medical terms. The translation of these medical terms must also be included in the translation. However, both our approach and existing approaches ignored the punctuations and non-Chinese characters in the translation candidate generation process; as a result, we only obtained the partial translation for those Chinese translations.

5.2 Discussion for rule based candidate generation with data mining approach

The rule based candidate generation with data mining approach has two major advantages. First, the rule based candidate generation is the first system available in

English to Chinese translation that handles Chinese translations which includes non-Chinese characters. Second, the data mining approach is not only able to select the correct translation candidates, but also generate a model that can be applied for OOV term translations.

The rule based candidate generation with data mining approach was able to reach a 95.48 % precision, however some translation candidates were still selected incorrectly by the decision tree. Since the decision tree tries to find a general rule for all pairs of English medical OOV term and its translation, regardless of frequencies of their appearances. Some translation pairs occur very infrequently are classified wrongly.

When using different features in the decision tree, association measures always play a very important role. However, there is a drawback from association measures as when we calculate the association measures, we obtained the numbers of returned pages from the Internet, and these numbers represent the co-occurrence of the input English OOV term and the Chinese translation candidates. These numbers do not represent the location distance between the input English OOV term and the Chinese translation candidates in the retrieved snippets. As a result, no matter how far or how close the Chinese translation candidate is away from the English OOV term, they all result the same numbers. Even we filtered out the translation pairs with far location distance; however few translation pairs pass the filter and cause the decision tree to classify wrongly. When a wrong translation pair being classified into a correct class, it maybe another disease whose Chinese translation co-occur together with the English OOV term more often than the correct one. For example (Tyrosinemial 酪胺基酸症) but the Chinese translation we got is “胺基酸代謝疾病” which is the translation of “Aminoacidopathies,” because “Tyrosinemial” appear together more often with the “胺基酸代謝疾病” than the correct one. Another type of error is the partial translation problem, for example (MLES MLES 症候群), our approach did not get the non-Chinese character part in that Chinese translation. The Chinese translation we got is “症候群,” because we matched the non-Chinese character part as the English OOV term. For both examples, they however have very high association measure scores, high frequency and not too far location distance. These results cause the decision tree to classify the incorrect translation pairs into a correct class.

When a correct translation pair being classified into a wrong class, it is usually caused by some Chinese word segmentation system in the search engine. For term (Myotubular Myopathy 肌小管病/肌小管病变), our retrieved answer is “肌小管病”, but when query this word back to Yahoo API to check for co-occurrence the result is zero, the co-occurrence for “肌小管病变” is more than 100. Same result with Google and MSN. So far, we can only assume that those search engines use a statically based Chinese word segmentation system, the system very much depends on the corpus that occurs on the Internet.

5.3 Compare the two approaches

Two common limitations in both of our approaches are firstly unable to find translation candidates in the process of querying OOV terms into the web; secondly some Chinese translations were not correct generated in the translation candidate generation process.

Some translations are lost in the process of web querying because there is no Chinese translation in text format for that specific English OOV term in the Internet from the retrieved snippets. This is because those Chinese translations may exist in a picture or PDF format.

For errors occurring in candidate generation process, most correct Chinese translation candidates can be found in the retrieved snippets when we manually check these snippets. But both of our approaches were unable to generate some of the correct Chinese translation candidates if that correct Chinese translation candidates occur very infrequently in the snippets. The precision of candidate generation for both of our approaches and existing approach are shown in Fig. 5.1. The overall precision for both of our approaches and existing approach are shown in Fig. 5.2.

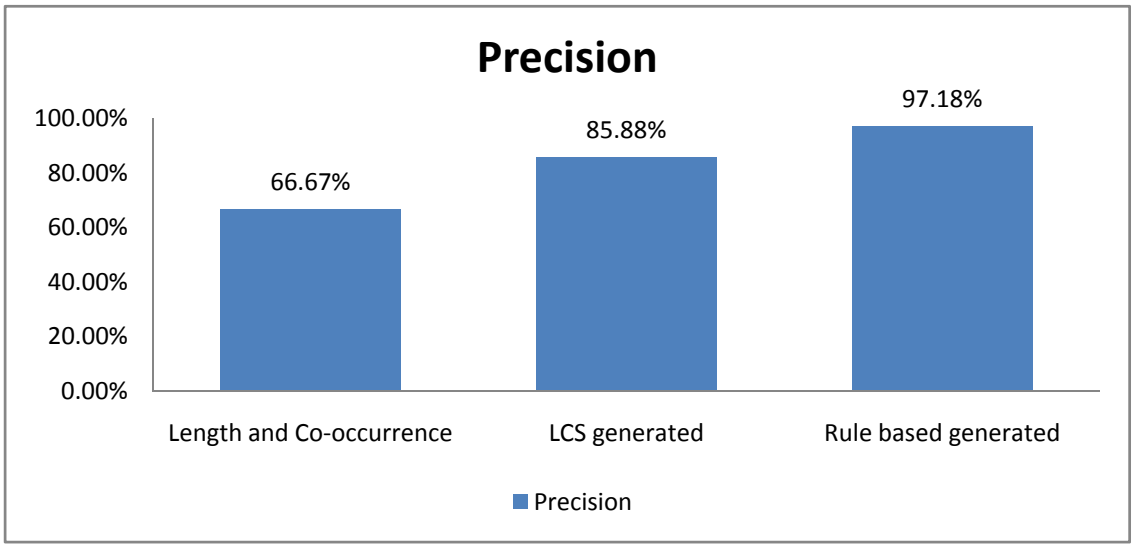

Figure 5.1 are the bar graphs of candidate generation precision of existing approach compare to both of our approaches.

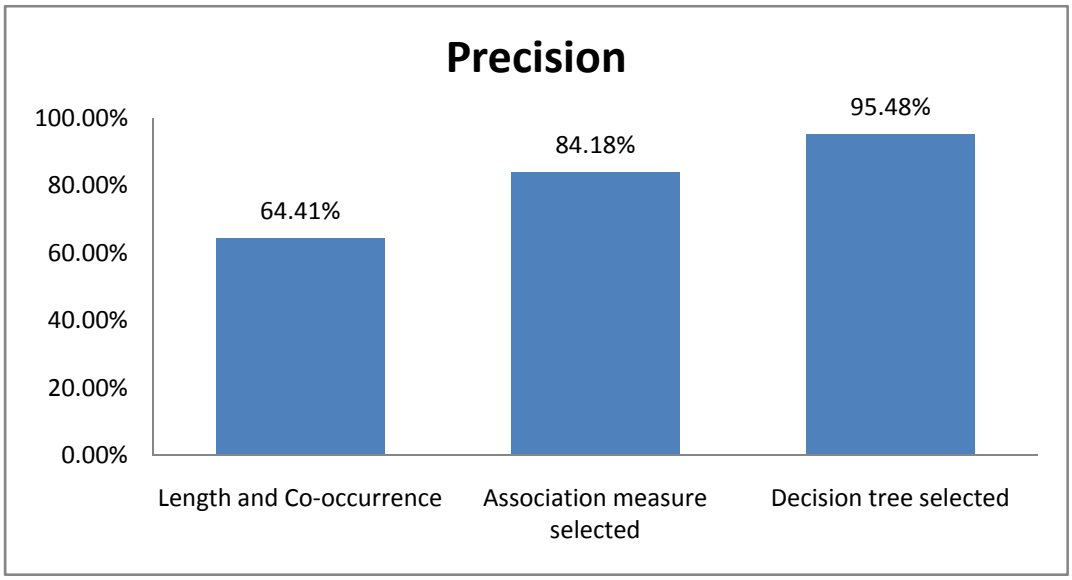

Figure 5.2 are the bar graphs of overall precision of existing approach compare to both of our approaches

If we simply compare experiment results for in Fig. 5.1 and 5.2, we can conclude that translation candidate selection in both approaches performed a very high precision. Most problems causing the overall performances to drop come from candidate generation process. The rule based candidate generation with data mining approach outperforms the LCS candidate generation with association measure approach, because rule based candidate generating approach generated a better set of translation candidate than LCS candidate generating approach. Improvements on translation candidates improves the overall performs. Our future work will majorly focus on how to generate better translation candidates.

Chapter 6

Conclusion

English-Chinese OOV translation extraction is a technique to automatically generate Chinese translations from the given English OOV terms. Many approaches have been proposed to achieve this task. However, the existing approaches usually require human to select the Chinese translation for the English OOV term from a list of possible candidates. And existing approaches focus only generating the Chinese translations with only Chinese characters. Yet some Chinese technical terms in the real world may compose of several kinds of characters. We proposed two approaches to solve the problems of the English-Chinese OOV term translation. We conduct the experiments based on ICD 9 English medical OOV terms. The first approach performs an automatic selection without human interfere and obtains an overall precision up to 84.18%. The second approach uses a rule based candidate generating system to solve the non-Chinese character problem. This new approach takes into account of each individual English OOV term and set different rules to generate the Chinese translation candidates. The selection process of translation candidates is done by decision tree, which generates rule based on all different features we extracted. The second approach outperforms the first approach with an overall precision of 95.48%; it also outperforms the length co-occurrences approach, which can only reach an overall precision of 64.41%. By compare both of our approaches in chapter 5; their performances are depending on translation candidate generating process. The better are the generated candidates, the more accurate is the result. Our future work will majorly focus on how to generate better translation candidates.

References

Agrawal, R., T. Imieli, et al. (1993). "Mining association rules between sets of items in large databases." SIGMOD Rec. 22(2): 207-216.

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

Ballesteros, L. and B. Croft (1996). Dictionary-based Methods for Cross-Lingual Information Retrieval. International Conference on Database and Expert Systems Applications: 791-801.

Bergroth, L., H. Hakonen, et al. (2000). A Survey of Longest Common Subsequence Algorithms. Proceedings of the Seventh International Symposium on String Processing Information Retrieval (SPIRE'00), IEEE Computer Society: 39.

Buitelaar, P., D. Olejnik, et al. (2004). A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis. The Semantic Web: Research and Applications, First European Semantic Web Symposium. Heraklion, Crete, Greece,. 3053/2004: 31-44.

Chen, H.-H., C. Yang, et al. (2003). Learning formulation and transformation rules for multilingual named entities. Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition - Volume 15. Sapporo, Japan, Association for Computational Linguistics: 1-8.

Cheng, K.-S., G. H. Young, et al. (1999). "A study on word-based and integral-bit Chinese text compression algorithms." J. Am. Soc. Inf. Sci. 50(3): 218-228.

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

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

Dan, G. (1997). Algorithms on Strings, Trees and Sequences:, Cambridge University Press.

DeDe. (2009). "Dede segmenation tool." from http://dedecms.com.

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

Government, C. (2009). "Chinese language." from http://www.china-language.gov.cn.

Gu, T., X. H. Wang, et al. (2004). An ontology-based context model in intelligent environments. In Communication Networks and Distributed Systems Modeling and Simulation Conference: 270-275.

Han, J. and M. Kamber (2006). Data Mining: Concepts and Techniques.

Hand, D., H. Mannila, et al. (2001). Principles of Data Mining. Massachusetts London, A Bradford Book The MIT Press Cambridge.

Health, B. o. (2009). "Key word list for medical terms."

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

Hull, D. A. and G. Grefenstette (1996). Querying across languages: a dictionary-based approach to multilingual information retrieval. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. Zurich, Switzerland, ACM: 49-57.

ICD9, C. (2009, 30 SEP 2009). "公告罕见疾病名單暨 ICD-9-CM 編碼一覽表." from http://web.cdc.gov.tw/public/Attachment/981111144376.pdf.

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

Jinshan. (2009). "Kingsoft Translation software." from www.kingsoft.com.

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

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

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

Language, C. (2009). "Languages of China." from http://www.chineselanguage.org.

Limungkura, T., T. Theeramunkong, et al. (2009). Extraction of Medical Expert-Affiliation Relatinos from WWW. The 6th International Joint Conference on Computer Science and Software Engineering. Puket, Thailand: 195-199.

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

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

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

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

Ngomo, A.-C. N. (2008). Knowledge-free discovery of domain-specific multiword units.

Proceedings of the 2008 ACM symposium on Applied computing. Fortaleza, Ceara, Brazil, ACM: 1561-1565.

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

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

Pei, J., J. Han, et al. (2001). Mining Frequent Itemsets with Convertible Constraints. Proceedings of the 17th International Conference on Data Engineering, IEEE Computer Society: 433.

Ponte, J. and W. Croft (1996). Useg: A Retargetable word segmentation procedure for information retrieval. Technical report:UM-CS-1996-002, University of Massachusetts.

Qu, J., K. Viriyayudhakorn, et al. (2009). Automatic English to Chinese Translation of Medical Terms using Association Rule Mining with Web Data. The 6th International Joint Conference on Computer Science and Software Engineering Phuket, Thailand: 336-341.

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

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

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

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

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

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

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

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

Viriyayudhakorn, K., T. Theeramunkong, et al. (2008). Mining Translation pairs for

Thai-English Medical Terms. KICSS, KICSS: 204-211.
WiKi, C. (2009). "Chinese language." from http://en.wikipedia.org/wiki/Chinese_language
Witten, I. H. and E. Frank (2002). "Data mining: practical machine learning tools and techniques with Java implementations." SIGMOD Rec. 31(1): 76-77.
Wu, Z. and G. Tseng (1993). "Chinese text segmentation for text retrieval: Achievements and problems." Journal of the American Society for Information Science and Technology 44(9): 532-542.
Yahoo. (2009). "Yahoo API." from http://developer.yahoo.com.
Yang, C. C. and K. W. Li (2003). "Automatic construction of English/Chinese parallel corpora." J. Am. Soc. Inf. Sci. Technol. 54(8): 730-742.
YE, N. and Q. CHEN (2001). "An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems." Quality and Reliability Engineering International 17(2): 105-112.
Yuan, Y. and M. J. Shaw (1995). "Induction of fuzzy decision trees." Fuzzy Sets and Systems Volume 69(Issue 2).
Zhang, T. (2000). Association Rules. Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications, Springer-Verlag: 245-256.
Zhang, Y., F. Huang, et al. (2005). Mining translations of OOV terms from the web through cross-lingual query expansion. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. Salvador, Brazil, ACM: 669-670.
Zhang, Y. and P. Vines (2004). Detection and translation of OOV terms prior to query time. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Sheffield, United Kingdom, ACM: 524-525.
Zhang, Y. and P. Vines (2004). Using the web for automated translation extraction in cross-language information retrieval. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Sheffield, United Kingdom, ACM: 162-169.
Zhang, Y., P. Vines, et al. (2005). "Chinese OOV translation and post-translation query expansion in chinese--english cross-lingual information retrieval." ACM Transactions on Asian Language Information Processing (TALIP) 4(2): 57-77.

Appendix A

Top 50 domains

#	Name of domains	Domains	Number of snippets
1	Taiwan foundation for rare disorders	www.tfrd.org.tw	324
2	Bureau of national health insurance, department of health	www.nhinb.gov.tw	319
3	Department of Health Bureau of Health advisory services to windows of genetic diseases	www.bhp-gc.tw	270
4	Yahoo knowledge	tw.knowledge.yahoo.com	257
5	Yahoo blog	tw.myblog.yahoo.com	234
6	hospital administration commission, department of health	www.cto.doh.gov.tw	160
7	taiwan human genetics society	www.genes-at-taiwan.com.tw	157
8	Department of health	www.doh.gov.tw	156
9	Lienchiang County Municipal Hospital	www.pngh.tpg.gov.tw	152
10	WRETCH AlbumFront Page	www.wretch.cc	138
11	Taipei Municipal Jianguo Senior High School	web.ck.tp.edu.tw	116
12	County Wenchang School	www.wcps.tcc.edu.tw	107
13	Chang Gung Memorial Hospital	www.cgmh.org.tw	105
14	Sibo Network - Chinese Electronic Periodical Services	www.ceps.com.tw	102
15	Taipei City Government Health Office	www.health.gov.tw	101
16	Kang-Ning General Hospital	www.knh.org.tw	95
17	National Cheng Kung University	www.ncku.edu.tw	94
18	Medscape	www.24drs.com	76
19	Pediatric Clinical Advisor	www.leaderbook.com.tw	73
20	Providence University	www.osa.pu.edu.tw	72
21	Bureau of National Health Insurance	www.nhitb.gov.tw	70
22	Disabled Services Information Network	disable.yam.org.tw	66
23	Taoyuan Hospital, Department of Health	www.tygh.gov.tw	58
24	Blog yam	blog.yam.com	51
25	*Wikipedia Chinese*	zh.wikipedia.org	48
26	National Taiwan University Hospital	ntuh.mc.ntu.edu.tw	47
27	Taipei Veterans General Hospital-Index	homepage.vghtpe.gov.tw	43
28	Mackay Memorial Hospital	www.mmh.org.tw	42
29	Buddhist Tzu Chi General Hospital	www.tzuchi.com.tw	40

30	Satoshi Kuni public hall	www.17885.com.tw	37
31	Special cases of a rare disease, a rare disease drugs flow cum Nutrition Center	www.rdfdlmc.tw	35
32	China Medical University Hospital	www.cmuh.org.tw	35
33	Health 99 Education Resource	health99.doh.gov.tw	35
34	Medical Kwiki	zh-tw.med.wikia.com	35
35	Newborn Screening. in *Taiwan*	nbs.tw	32
36	CHI MEI MEDICAL CENTER	www.chimei.org.tw	30
37	Kaohsiung Medical University	www.kmu.edu.tw	30
38	Health and Medical Information	tw16.net	28
39	National Yang-Ming University	www.ym.edu.tw	27
40	union News	udn.com	26
41	The new family of rare diseases Paradise	www.rare.org.tw	25
42	Kaohsiung Medical University Chung-Ho Memorial Hospital	www.kmuh.org.tw	25
43	Blog Sina	blog.sina.com.tw	24
44	Chang Gung University	www.cgu.edu.tw	23
45	*MEDgle* medical and health search	www.medgle.com.tw	23
46	China Medical University , Graduate Institute of Basic Medical	www2.cmu.edu.tw	22
47	Epoch Times News Network	epochtimes.com	22
48	National Institute of Neurological and Allied Sciences	www.neuro.org.tw	21
49	myblog yahoo	hk.myblog.yahoo.com	21
50	The New England Regional Genetics Group	www.nergg.org	20

Appendix B

ICD 9 Table

ID	English ICD9 medical terms	Correction Chinese translations	Our Chinese translation
1	Urea cycle disorders	尿素循环代谢障碍	尿素循環代謝障礙
2	Citrullinemia	瓜胺酸血症	瓜胺酸血症
3	Amino acid metabolic disorders	胺基酸代谢疾病	胺基酸代謝疾病
4	Aminoacidopathies	胺基酸代谢疾病	胺基酸代謝疾病
5	Homocystinuria	高胱胺酸尿症	高胱胺酸尿症
6	Hypermethioninemia	高甲硫胺酸血症	高甲硫胺酸血症
7	Cystinosis	胱胺酸症	胱胺酸症
8	Nonketotic hyperglycinemia	非酮性高甘胺酸血症	非酮性高甘胺酸血症
9	Phenylketonuria	苯酮尿症	苯酮尿症
10	Tetrahydrobiopterin deficiency	四氢基喋呤缺乏症	四氫基喋呤缺乏
11	Hereditary tyrosinemia	遗传性高酪胺酸血症	遺傳性高酪胺酸血症
12	Maple syrup urine disease	枫糖尿症	楓糖尿症
13	porphyria	紫质症	紫質症
14	Multiple sclerosis	多发性硬化症	多發性硬化症
15	Gaucher's disease	高雪氏症	高雪氏症
16	Wilson's disease	威尔森氏症	威爾森氏症

17	Nesidioblastosis	胰岛母细胞瘤	胰島母細胞瘤
18	Persistent hyperinsulinemic hypoglycemia of infancy	持續性幼兒型胰島素過度分泌低血糖症	持續性幼兒型胰島素過度分泌低血糖症
19	Amyotrophic lateral sclerosis	肌萎缩性侧索硬化症	肌萎縮性脊髓側索硬化症
20	Organic acidemias	有机酸血症	有機酸血症
21	lsovaleric acidemia	异戊酸血症	異戊酸血症
22	Glutaric aciduria type	戊二酸血症，第一、二型	戊二酸血症
23	Propionic acidemia	丙酸血症	丙酸血症
24	Methylmalonic acidemia	甲基丙二酸血症	甲基丙二酸血症
25	3-Hydroxy-3-methyl-glutaric acidemia	3-氢基-3-甲基戊二酸血症	3-氫基-3-甲基戊二酸血症
26	Galactosemia	半乳糖血症	半乳糖血症
27	Fatty acid oxidation defect	脂肪酸氧化作用缺陷	脂肪酸氧化作用缺陷
28	Carnitine deficiency syndrome	原发性肉碱缺乏症	原發性肉鹼缺乏
29	Mitochondrial defect	粒腺体缺陷	粒線體缺陷
30	Kearns Sayre syndrome	Kearns Sayre 氏症候群	Kearns Sayre 氏症候群
31	Leigh disease	Leigh 氏童年期脑脊髓病变	Leigh 氏童年期腦脊髓病
32	MELAS	MELAS 症候群	症候群
33	Mitochondrial Neurogastrointestinal Encephalopathy Syndrome	MNGIE 症候群粒腺体性神经胃肠脑病变症候群	症候群粒線體性神經胃腸腦病變症候群
34	Aarskog-Scott syndrome	Aarskog-Scott 氏症候群	Aarskog-Scott 氏症候群
35	Achondroplasia	软骨发育不全症	軟骨發育不全症

36	Angelman syndrome	Angelman 氏症候群	Angelman 氏症候群
37	Ataxia telangiectasia	共济失调微血管扩张症候群	共濟失調微血管擴張症候群
38	Cockayne syndrome	Cockayne 氏症候群	Cockayne 氏症候群
39	Duchenne muscular dystrophy	裘馨氏肌肉失养症	裘馨氏肌肉失養症
40	Glycogen storage disease	肝醣储积症	肝醣儲積症
41	GM1/GM2 gangliosidosis	GM1/GM2 神经节苷脂储积症	GM1/GM2 神經節甘脂儲積症
42	Hereditary epidermolysis bullosa	遗传性表皮分解性水疱症	遺傳性表皮分解性水皰症
43	Huntington disease	亨汀顿氏舞蹈症	亨丁頓舞蹈症
44	Huntington's chorea	亨汀顿氏舞蹈症	亨汀頓氏舞蹈症
45	Hutchinson Gilford progeria syndrome	早老症	早老症
46	lchthyosis	层状鱼鳞癣（自体隐性遗传型）	層狀魚鱗癬
47	lamellar recessive	层状鱼鳞癣（自体隐性遗传型）	層狀魚鱗癬
48	Kenny-Caffey syndrome	Kenny-Caffey 氏症候群	Kenny-Caffey 氏症候群
49	Lesch-Nyhan syndrome	Lesch-Nyhan 氏症候群	Lesch-Nyhan 氏症候群
50	Lowe sydrome	Lowe 氏症候群	Lowe 氏症候群
51	Mucopolysaccharidoses	黏多醣症	黏多醣症
52	Osteogenesis imperfecta	成骨不全症	成骨不全症

53	Pseudohypoparathyroidism	假性副甲状腺低能症	假性副甲狀腺低能症
54	Rett syndrome	瑞特氏症候群	瑞特氏症候群
55	Spinal muscular atrophy	脊髓性肌肉萎缩症	脊髓性肌肉萎縮症
56	Spinocerebellar ataxia	脊髓小脑性共济失调	小腦萎縮症
57	Sulfite oxidase deficiency	亚硫酸盐氧化酶缺乏	亞硫酸鹽氧化酶缺乏
58	Thalassemia major	重型海洋性贫血	重型海洋性貧血
59	Tuberous sclerosis	结节性硬化症	結節性硬化症
60	Waardenburg syndrome	瓦登伯格氏症候群	瓦登伯格氏症候群
61	X-linked hypophosphatemic rickets	性连遗传型低磷酸盐佝偻症	性連遺傳型低磷酸鹽佝僂症
62	Zellweger syndrome	Zellweger 氏症候群	Zellweger 氏症候群
63	Progressive intrahepatic cholestasis	进行性家族性肝内胆汁滞留症	進行性家族性肝內膽汁滯留症
64	Inbon errors of bile acid synthesis	先天性胆酸合成障碍	先天性膽酸合成障礙
65	Primary Paget disease	原发性变形性骨炎	原發性變形性骨炎
66	Nitroacetylglutamate synthetase deficiency	乙酰穀胺酸合成酶缺乏症	乙醯穀胺酸合成酶缺乏症
67	NAG synthetase deficiency	乙酰穀胺酸合成酶缺乏症	乙醯穀胺酸合成酶缺乏症
68	Ornithine transcarbamylase deficiency	鸟胺酸氨甲酰基转移酶缺乏症	鳥胺酸氨甲醯基轉移酶缺乏症
69	Apert syndrome	爱伯特氏症	愛伯特氏症

70	Cleidocranial dysplasia	锁骨颅骨发育异常	鎖骨顱骨發育異常
71	DiGeorge's syndrome	DiGeorge's 症候群	DiGeorge's 症候群
72	Homozygous familial hypercholesterolemia	同合子家族性高胆固醇血症	同合子家族性高膽固醇血症
73	Fucosidosis	岩藻糖代谢异常（储积症）	岩藻糖代謝異常
74	PAH type PKU combine with Sucrase-isomaltase deficiency	典型苯酮尿症合并蔗糖酶同麦芽糖酶缺乏症	典型苯酮尿症合併蔗糖同麥芽糖缺乏症
75	Nemaline Rod Myopathy	Nemaline 线状肌肉病变	Nemaline 線狀肌肉病
76	Fibrodysplasia Ossificans Progressiva	进行性骨化性肌炎	進行性骨化性肌炎
77	Menkes syndrome	Menkes 氏症候群	Menkes 氏症候群
78	Fabry disease	Fabry 氏症	法布瑞氏症
79	Prader-Willi syndrome	Prader-Willi 氏症候群	威利症
80	Niemann-Pick disease	Niemann-Pick 氏症，鞘髓磷脂储积症	尼曼匹克症
81	Tricho-hepato-enteric syndrome	发-肝-肠症候群	髮-肝-腸症候群
82	Collodion baby	胶膜儿	膠膜兒
83	Harlequin ichthyosis	斑色鱼鳞癣	斑色魚鱗癬
84	Bullous Congenital ichthyosiform erythoderma	水泡型先天性鱼鳞癣样红皮症（表皮松解性角化过度症）	水泡型先天性魚鱗癬樣紅皮症
85	epidermolytic hyperkeratosis	層狀魚鱗癣	層狀魚鱗癬
86	Laron syndrome	Laron 氏侏儒症候群	氏侏儒症候群

87	Laron Dwarfism	Laron 氏侏儒症候群	Laron 氏侏儒症候群
88	Smith-Lemli-Opitz syndrome	Smith-Lemli-Opitz 氏症候群	Smith-Lemli-Opitz 氏症候群
89	Bardet-Biedl syndrome	Bardet-Biedl 氏症候群	巴德-畢德氏症候群
90	Larsen syndrome	Larsen 氏症候群, 顎裂-先天性脱位症候群	Larsen 氏症候群
91	Sialidosis	涎酸酵素缺乏症	涎酸酵素缺乏症
92	Alstrom Syndrome	Alstrom 氏症候群	Alstrom 氏症候群
93	Chronic primary granulomatous disease	原发性慢性肉芽肿病	原發性慢性肉芽腫病
94	Persistent hyperinsulinemic hypoglycemia of infancy PHHI	持续性幼儿型胰岛素过度分泌低血糖症	持续性幼儿型胰岛素过度分泌低血糖症
95	Familial hyperchylomicronemia	家族性高乳糜微粒血症	家族性高乳糜微粒血症
96	W A G R syndrome	威尔姆氏肿瘤、无虹膜、性器异常、智能障碍症候群（W A G R 症候群）	Lost
97	Wilms' tumor-Aniridia-Genitourinary Anomalies-mental Retardation	威尔姆氏肿瘤、无虹膜、性器异常、智能障碍症候群（W A G R 症候群）	Lost
98	Ectodermal Dysplasias	外胚层增生不良症	外胚層增生不良症
99	Beckwith Wiedemann syndrome	Beckwith Wiedemann 氏症候群	Beckwith Wiedemann 氏症候群

100	Congenital insensitivity to pain with anhidrosis	先天性痛不敏感症合并无汗症	先天性痛不敏感症合併無汗症
101	Wolfram syndrome	Wolfram 氏症候群	Wolfram 氏症候群
102	Adrenoleukodystrophy	肾上腺脑白质失养症	腎上腺腦白質失養症
103	McCune Albright syndrome	McCune Albright 氏症候群	McCune Albright 氏症候群
104	Crouzon syndrome	Crouzon 氏症候群	Crouzon 氏症候群
105	Thrombasthenia	血小板无力症	血小板無力症
106	Schwartz Jampel syndrome	Schwartz Jampel 氏症候群	Schwartz Jampel 氏症候群
107	Fraser syndrome	Fraser 氏症候群	Fraser 氏症候群
108	Mucolipidosis	黏脂质症	黏脂質症
109	Ehlers Danlos syndrome	先天结缔组织异常第四型	先天結締組織異常
110	Myotonic dystrophy	肌肉强直症	肌肉強直症
111	Congenital Hyper IgE syndrome	先天性高免疫球蛋白 E 症候群	先天性高免疫球蛋白
112	Tyrosinemial	酪胺基酸症第一型、第二型、第三型	酸代謝疾病
113	Hereditary tyrosinemia	酪胺基酸症第一型、第二型、第三型	代謝疾病
114	Hyperlysinemia	高离氨基酸血症	高離氨基酸血症
115	Histidinemia	组胺酸血症	組胺酸血症

116	3-Methylcrotonyl-CoA carboxylase deficiency	三甲基巴豆酰辅酶 A 化酵素缺乏症	三甲基巴豆醯輔酶 A 梭化酵素缺乏症
117	Multiple carboxylase deficiency	多發性梭化酶缺乏症	多發性梭化酶缺乏症
118	Split-hand/ Split-foot malformation	裂手裂足症	裂手裂足症
119	Metachromatic Leukodystrophy	MLD 症候群	症候群
120	Campomelic dysplasia with autosomal sex reversal	短指发育不良及性别颠倒	短指發育不良及性別顛倒
121	Osteopetrosis	骨质石化症	骨質石化病
122	Carbohydrate-deficiency glycoprotein syndrome	碳水化合缺乏醣蛋白症候群	碳水化合缺乏醣蛋白症候群
123	Trimethylaminuria	臭鱼症	臭魚症
124	Congenital generalized lipodystrophy	先天性全身脂质营养不良症	先天性全身脂質營養不良症
125	Multiple pterygium syndrome	多发性翼状膜症候群	多發性翼狀膜症候群
126	Idiopathic Infantile Arterial Calcification	特发性婴儿动脉硬化症	特發性嬰兒動脈硬化症
127	Miller Dieker syndrome	Miller Dieker 症候群	Miller Dieker 症候群
128	Medium-chain acyl-coenzyme A dehydrogenase deficiency	中链脂肪酸去氢酵素缺乏症	中鏈脂肪酸去氫酵素缺乏症
129	Hyperprolinemia	高脯胺酸血症	高脯胺酸血症
130	Cystic fibrosis	囊状纤维化症	囊性纖維化
131	Pyruvate dehydrogenase deficiency	丙酮酸盐脱氢酶缺乏症	丙酮酸鹽脫氫酶缺乏症
132	Neuronal ceroid lipofuscinosis	神经元蜡样脂褐质储积症	神經元蠟樣脂褐質儲積症
133	Meleda disease	Meleda 岛病	Meleda 島病

134	Neurofibromatosis	神经纤维瘤症候群第二型	神經纖維瘤
135	Alexander disease	Alexander 氏病	亞歷山大症
136	ACTH resistance	肾上腺皮促素抗性	腎上腺皮促素抗性
137	1α-hydroxylase deficiency	1α-羟化酶缺乏症候群	1α-羥化酶缺乏症候群
138	Stiffperson syndrome	僵体症候群	僵體症候群
139	Primary Pulmonary Hypertension	原发性肺动脉高压	原發性肺動脈高壓症
140	Cornelia de Lange syndrome	Cornelia de Lange 氏症候群	Cornelia de Lange 氏症候群
141	Pseudoachondroplastic dysplasia	假性软骨发育不全	假性軟骨發育不全
142	Rubinstein-Taybi syndrome	Rubinstein-Taybi 氏症候群	Rubinstein-Taybi 氏症候群
143	Facioscapulohumeral muscular dystrophy	面肩胛肱肌失养症	面肩胛肱肌失養症
144	Bartter's syndrome	Bartter 氏症候群	巴特氏症候群
145	Short-chain acyl-CoA dehydrogenase deficiency	短链脂肪酸去氢酶缺乏症	短鏈脂肪酸去氫酶缺乏症
146	Homozygous proetin C deficiency	同基因合子蛋白质 C 缺乏症	同基因合子蛋白質 C 缺乏症
147	α1- Antitrypsin deficiency	α1-抗胰蛋白酶缺乏症	α1-抗胰蛋白酶缺乏症
148	Tyrosine hydroxylase deficiency	酪胺酸羟化酶缺乏症	酪胺酸羥化酶缺乏症
149	Aromatic L-amino acid decarboxylase deficiency	芳香族 L-胺基酸类脱羧基酶缺乏症	芳香族 L-胺基酸類脫羧基酶缺乏症
150	Congenital adrenal hypoplasia	先天性肾上腺发育不全	先天性腎上腺發育不全

151	Kallmann syndrome	Kallmann 氏症候群	Kallmann 氏症候群
152	Bruton's agammaglobulinemia	布鲁顿氏低免疫球蛋白血症	布魯頓氏低免疫球蛋白血症
153	Wiskott-Aldrich Syndrome	Wiskott-Aldrich 氏症候群	Wiskott-Aldrich 氏症候群
154	Severe combined immunodeficiency	严重复合型免疫缺乏症	嚴重複合型免疫缺乏症
155	Complement Component 8 deficiency	补体成份 8 缺乏症	補體成份 8 缺乏症
156	Holt-Oram Syndrome	Holt-Oram 氏症候群	Holt-Oram 氏症候群
157	Hereditary spastic paraplegia	遗传性痉挛性下身麻痹	遺傳性痙攣性下身麻痺
158	IPEX Syndrome	IPEX 症候群	IPEX 症候群
159	Williams Syndrome	威廉斯氏症候群	威廉氏症
160	Joubert syndrome	Joubert 氏症候群, 家族性小脑蚓部发育不全	Joubert 氏症候群
161	Congenital Interstitial Cell of Cajal Hyperplasia with Neuronal Intestinal Dysplasia	先天性 Cajal 氏间质细胞增生合并肠道神经元发育异常	Lost
162	Hallerman-Streiff Syndrome	海勒曼-史德莱夫氏症候群	海勒曼-史德莱夫氏症候群
163	Kabuki syndrome	歌舞伎症候群	歌舞伎症候群
164	Oto-Palato-Digital syndrome	耳-腭-指（趾）症候群	氏症候群
165	Pelizaeus-Merzbacher Disease	Pelizaeus-Merzbacher 氏症, 慢性儿童型脑硬化症	Pelizaeus-Merzbacher 氏症

166	Charcot Marie Tooth Disease	Charcot Maire Tooth 氏症，进行性神经性腓骨萎缩症	Charcot Marie Tooth 氏症
167	Cerebrotendinous Xanthomatosis	脑腱性黄瘤症	腦腱性黃瘤症
168	Darier's disease	Darier 氏症，毛囊角化病	Darier 氏症
169	Conradi-Hunermann syndrome	Conradi-Hunermann 氏症候群	Conradi-Hunermann 氏症候群
170	Andersen syndrome	Andersen 氏症候群（心节律障碍暨周期性麻痹症候群；钾离子通道病变）	Andersen 氏症候群
171	Kennedy Disease	肯尼迪氏症(脊髓延髓性肌肉萎缩症)	甘迺迪氏症
172	Dyskeratosis Congenita	先天性角化不全症	先天性角化不全症
173	Treacher Collins Syndrome	Treacher Collins 氏症候群	Treacher Collins 氏症候群
174	Familial Amyloidotic Polyneuropathy	家族性淀粉样多发性神经病变	家族性澱粉樣多發性神經病
175	Robinow Syndrome	Robinow 氏症候群	Robinow 氏症
176	Hereditary Hemorrhagic Telangiectasia	遗传性出血性血管扩张症	遺傳性出血性血管擴張症
177	Glut1 deficiency syndrome	脑血管屏障葡萄糖输送缺陷	Lost
178	Glucose Transporter deficiency syndrome	脑血管屏障葡萄糖输送缺陷	Lost

179	Hyperornithinemia-Hyperammonemia-Homocitrullinuria Syndrome	高鸟胺酸血症-高氨血症-高瓜胺酸血症症候群	高鳥胺酸血症-高氨血症-高瓜胺酸血症
180	Myotubular Myopathy	肌小管病变	肌小管病
181	Pfeiffer syndrome	Pfeiffer 氏症候群	Pfeiffer 氏症候群
182	Pantothenate Kinase Associated Neurodegeneration	泛酸鹽激酶關聯之神經退化性疾病	泛酸鹽激酶關聯之神經退化性疾病
183	Hyper-IgM syndrome	高免疫球蛋白 M 症候群	高免疫球蛋白 M 症候群
184	Nail-Patella Syndrome	指（趾）甲髕骨症候群	症候群

Appendix C

Curriculum Vitae

Education:

M.S.IT in Information Science, Thammasat University, Thailand
(November 2007- Present)

B.BA in Business administration and finance, Summa Cum Laude honor, Ramkhamhaeng University, Thailand
(June 2003-November 2006)

Publications:

National Journal

- 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. (Conditional Accepted, revised and under re-reviewing process).

International Conferences

- 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. Joint Conference on Computer Science and Software Engineering, 336-341. Phuket, Thailand, May 2009.

- Jian Qu, Thanaruk Theeramunkong, Cholwich Nattee and Pakinee Aimmanee. A Novel Candidate Generation Technique for Web based English- Chinese Medical OOV Term Translation. International Conference on Knowledge, Information and Creativity Support Systems, 61-68. Seoul, Korea, Nov 2009. Published by JAIST.