April 2015

Automated Building of Classic Chinese-English Dictionary and Chinese-Hungarian Dictionary

Hongbo Fang
Worcester Polytechnic Institute

Xiaosong Wen
Worcester Polytechnic Institute

WORCESTER POLYTECHNIC INSTITUTE

Automated Building of Classic Chinese-English Dictionary and Chinese-Hungarian Dictionary

A Major Qualifying Project Report

submitted to the faculty of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements

for the Degree of Bachelor of Science by

Fang, Hongbo

Wen , Xiaosong

27 April, 2015

Professor Gábor N. Sárközy, Major Advisor

Abstract

Although there have been a great number of achievements in Natural Language Processing, there are still few studies available for languages between a character-based language and a word-based language, due to fundamental linguistic differences. This paper describes a methodology to build a bilingual Hungarian-Chinese dictionary and classic Chinese-English dictionary. This project is a sub-project of the Sztaki Dictionary project [5]. This report addresses the required Background, Methodology, Results and Evaluation, Conclusion, and some future work.

Key words: sentence-level parallel corpus, automatic bilingual dictionary, PMI, ancient language

Acknowledgments

This project is supported by

- Worcester Polytechnic Institute (WPI)
- Computer and Automation Research Institute, Hungarian Academy of Sciences (MTA-SZTAKI)

We would like to thank WPI MQP Advisor **Gábor Sárközy** and MTA-SZTAKI Project Advisor **András Kornai** for their valuable support and guidance.

Table of Contents

Tables and Figures:

Summary

In this project, we automatically built a parallel corpus and a bilingual dictionary between Classic Chinese and Hungarian. We evaluated and improved a Modern Chinese (Simplified) and Hungarian dictionary.

Chapter 1 describes the background, related work, and some mathematical concept that we used to analyze the dictionary. Chapter 2 explains the detailed methodology and techniques that we used in this project. Chapter 3 presents our results and evaluations. Chapter 4 concludes this project and Chapter 5 discusses the potential future work. Appendix A describes the tools we created in this project. Appendix B is a simple user guide for creating one's own parallel corpus and bilingual dictionary in four steps using our methodology. Appendix A describes the programs we created in this project.

CHAPTER 1: BACKGROUND AND INTRODUCTION

Before attempting to go through the whole project, it was necessary and important for us to do some research to be more familiar with the topic. In this chapter, we are introducing the history of the theory, the basic techniques and the mathematical methods we used to evaluate and analyze our dictionary.

1.1. Bilingual Dictionary

Bilingual Dictionary is a specialized dictionary used to translate words or phrases from one language to another. Bilingual dictionaries can be either unidirectional (list the meanings of words of one language in another) or bidirectional (allowing translation to and from both languages).

Not only does the bilingual dictionary play an important role in Natural Language Processing, widely applied in machine translation, but it also helps in many other fields including cross-language information retrieval [18] and cross-language plagiarism detection [22]. Early researchers used a context-based approach to create automatic bilingual dictionaries [14]

In this project, we used a context-based approach to create a bilingual Classic Chinese – English dictionary. There are two reasons why we created our own Chinese-Hungarian dictionary. First, we failed to find available Chinese-Hungarian dictionaries online that have both good quality and good coverage. Second, creating automatic bilingual dictionaries is an important problem for machine translations between medium-density languages (Chinese) and low-density languages (Hungarian).

1.2. Modern Chinese and Classic Chinese

1.2.1 The transition of Chinese

Chinese is a language that is widely used by a huge population with many varieties which has led to lots of other languages. Tracing from ancient generations thousands of years back, both spoken and written Chinese have evolved significantly throughout the 2,500 years of its history. Chinese has different compositions including Hieroglyphics, Semasiographic, and Phonograms.

Classic Chinese can be divided into two forms. One style is called *Ancient Chinese*. It has a written style originated from the spoken language before the Qin Dynasty and it also appears in the articles written by later generations. The so-called ancient language that specialists study nowadays is actually Ancient Chinese, which is very widely used by many poets and writers such as Han Yu, Liu Zongyuan, etc.

Another style form is called the *Ancient Vernacular* (白话文), starting from the six Dynasties (222-589) that is based upon Northern dialects. However, comparing with the language people utilized during the 20th century and later, Ancient Vernacular did not show much difference from modern Chinese. Although some ancient words still remain in Modern Chinese, we are still able to understand its meanings if we fully understand Ancient Chinese.

We will study more Ancient Chinese in this project report. Moreover, Ancient Chinese varies and appears differently in different Dynasties, thus our project will focus on the works before the Qin Dynasty such as *The Analects*', '*Book of Poetry*', and '*Art of War, etc.*

1.2.2 The difference between Modern Chinese and Classic (Ancient) Chinese

There are tremendous differences between Modern Chinese (both simplified Chinese and traditional Chinese) and Classic Chinese. Traditional Chinese still keeps most of the written style that originally came from Classic Chinese; however, in Modern Chinese, it has been totally simplified and reformed.

The way words are formed is very different as well. In Classic Chinese, every single non-functional character has its unique meaning, which is where the meaning of the word originally comes from. However in Modern Chinese, a word has only one meaning when characters are put together. For instance, in Classic Chinese, '妻子' can be separated into '妻' and '子', which has individual meanings as wife and son. Differently, in Modern Chinese, '妻子' only indicates wife as a uniformed word.

Moreover, grammar and phrases have evolved dramatically, which made the transition of Chinese significant in history. Plenty of special phrases were used in Ancient Chinese but no longer exist in Modern Chinese, so this makes it harder for people in the later generation to translate word by word. Also, some of the Chinese characters shared a similar meaning or pronunciation with either the name of the Emperors or other name-related terms, so that earlier generations tried to avoid using those words due to the existence of taboo.

Consequently, the differences between all forms of Chinese show the evolution and diversity of this amazing ancient language throughout history. Studying ancient Chinese has a historical significance toward human world

1.3. Parallel Corpus in Natural Language Processing

1.3.1 Machine Translation and Parallel Corpus

Due to economic globalization and information explosion, almost every corner of the world has a great number of people who are not native-speakers. In many scenarios, people tend to understand or have to use the language that they do not speak. However, according to Grime [9], today's world is facing the challenge of uneven distribution: several large languages, such as English and Chinese, account for 40% of the population; on the contrary, over 5000 small languages account for only about 4% of the population. This unbalanced distribution really makes people, who speak those small languages, difficult to get information in the languages they do not speak. All these factors made machine translation a very active research field in natural language processing.

A major approach to machine translation researchers are using today is using a parallel corpus to develop mathematical models. Parallel corpus is a text placed alongside its translation or translations, defined by Wikipedia (http://en.wikipedia.org/wiki/Parallel_text) [20]. The parallel corpus can be divided into two categories. One is *Comparable corpus*: the texts are of the same kind and cover the same content. An example would be different Wikipedia pages for sunglasses in different languages. Although they are not the exact translations, they are still talking about the same object. Another one is *Translation corpus*: the texts in one language (Hungarian) are translations of texts in the other language (Chinese) [24]. In the first part of our project, we were using huge quantities of translation corpuses, with size of more than one gigabyte, to recursively build a dictionary. After getting a dictionary with considerable quality,

we tested the dictionary on a Comparable corpus which is all Wikipedia pages downloaded from internet.

Therefore, the quality of the parallel corpus plays a really important role for the mathematical model. To build a high quality dictionary, corpuses with high word coverage and accurate alignment are highly required. We use a lot of collecting and processing techniques to improve both quality and quantity of the parallel corpus.

1.3.2 Techniques for Collecting and Processing Parallel Corpus

There are two main techniques for collecting parallel corpus: web mining, an automatic downloading method, and manual collection.

Web mining is a technique of letting people download and extract information from the internet. This tool has already helped people save a lot of labor. There are numerous projects using this tool. For example, STRAND [23] is one of the earliest parallel corpora created by web mining. There are plentiful open resources for parallel corpus: Wikipedia, the most popular and largest general reference, and cuyoo.com, a website with parallel Chinese and English.

Manual Collection is the most classic way to collect data. Although, with a cost of a great number of human labor, manual collection will provide a better quality and diversity and better language coverage. One of the good examples is the Chinese Text Project (http://ctext.org/)[3]. Containing over ten thousand titles and more than one billion characters, the Chinese Text Project is one of the largest databases of pre-modern Chinese texts in existence.

Meanwhile, there are some open-source projects we could use. One good project we used is UM-Corpus [16] and Hunglish [2]. UM-Corpus has both good quality and varieties of content. The data included embraces eight different domains: *News, Spoken, Laws Thesis, Educational*

Materials, Science, Subtitles and Microblog. Hunglish is a free, sentence-aligned Hungarian-English parallel corpus of about 120 million words in 4 million sentence pairs.

1.3.3 Parallel Corpus between Chinese and Hungarian – Languages with Fundamental Linguistic Differences

One of the most significant challenges we were facing in this project was the linguistic differences between Hungarian and Chinese.

Firstly, which also was the most significant difference, the basic units of composition to the two languages are different. In Chinese, every word consists of one or more than one characters. Meanwhile, each character has its own multiple meanings. In some situation, the meanings of the characters in a word make the meaning of that word. However, in other cases, the word has totally different meanings from the meaning of its components. In Hungarian, the basic unit of a word is a letter which is in a known set: alphabet. This is the same as most of the Western languages, for example: English, German, French, etc.

Secondly, the grammars of the two languages are different. In Hungarian, again, like English, there are different forms for a noun, for instance: *egy ember for* 'a person', and *emberek* for 'people'. However, there are no plural forms of nouns and no time tenses for a verb in Chinese. Chinese tends to rely on context to indicate that something happened in the past, using some time words or functional words to present the tense [8].

Last but not least, the word segmentation in Chinese is much harder than it is in Hungarian. In Hungarian, there is a specific symbol space for word segmentation. However, there is no word delimiter between words in written Chinese sentences and the length of each

word is very short [15]. In fact, there were even no punctuation marks in Chinese until the 20th century.

These major differences in linguistics make the project much harder, compared to build a dictionary between two western languages. Apparently, it is important and essential to apply different methods to Hungarian and Chinese before processing the alignment, especially when doing world-level and sentence-level alignments. Extra methods should also be used for the dictionary precision such as stemming.

1.4 Mathematical Concepts to Analyze Dictionary

To build a dictionary and to improve the quality of the dictionary, we used a parallel corpus aligned by a previous MQP group [29]. The size of parallel corpus was more than one gigabytes, including 8,582,791 Hungarian words and 1,782,380 Chinese characters. Meanwhile, the dictionary we built had more than twenty thousand pairs. Thus there was no way we could analyze and evaluate the dictionary manually. As a result, we used some numerical analysis methods to help us deal with such a great amount of data. It is essential to understand some concepts ahead.

1.4.1 Sparse Matrix

In numerical analysis, a sparse matrix is a matrix in which most of the elements are zero [7]. The fraction of zero elements over the total number of elements in a matrix is called the sparsity. Figure 1 is an example of a sparse matrix. Large sparse matrices are usually applied in scientific or engineering applications when solving partial differential equations. A huge collection of sparse matrix has been done by the University of Florida (2011) [26]. The

Collection could be widely used by the numerical linear algebra computations for the development and performance evaluation of sparse matrix algorithms. Because performance results with artificially generated matrices can be misleading and matrices are curated and made publicly available in many formats, the sparse matrix is allowed for robust and repeatable experiments. Sparse matrix could also be used in computer graphics applications to perform high-intensity numerical simulation on GPU [13]. The usage of a sparse matrix in this project will be discussed in section 2.2.1.

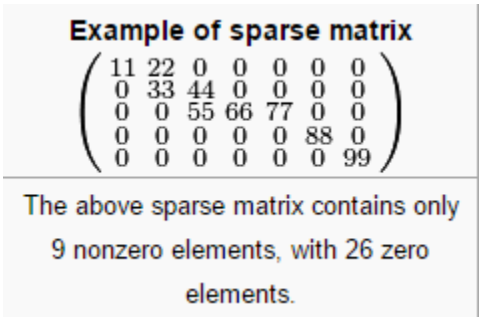

Figure 1: Example of a sparse matrix

1.4.2 Conditional Probability

Conditional probability measures the probability of an event given that (by assumption, presumption, assertion or evidence) another event has occurred [10]. If the event of interest is A and the event B is known or assumed to have occurred, 'the conditional probability of A given B', or 'the probability of A under the condition B', is usually written as P(A|B), or sometimes $P_B(A)$. If P (A|B) = P (A), A and B are considered to be independent.

In our project, conditional probability is used in pointwise mutual information (in Section 1.4.3 and Section 2.2.3)

1.4.3 Pointwise mutual information (PMI)

Pointwise mutual information, also known as point mutual information is a measure of association used in information theory and statistics [4][11]. In information theory, mutual information is more often defined as holding between random variables. Fano [6] originally defined mutual information between particular events X and Y, in our project the appearances of two particular words or characters.

Assuming two randomly distributed variables X and Y are independent, PMI shows the discrepancy between the probability of their joint distribution and their individual distributions. Mathematically:

$$PMI(X,Y) = \log\frac{p(X,Y)}{p(X)p(Y)} = \log\frac{p(X|Y)}{p(X)} = \log\frac{p(Y|X)}{p(Y)}$$

In the formula above, $p(X,Y)$ means the probability that both X and Y occurs, $p(X)$ is the probability that X occurs; $p(Y)$ is the probability that Y occurs. $p(X|Y)$ means the conditional probability that X occurs under the condition that Y already occurred and $p(Y|X)$ means the conditional probability that Y occurs under the condition that X already occurred

In this project, we used PMI to analyze our sparse matrix. The bigger the PMI is, the more likely the word pairs considered to be a pair. This will be discussed further in section 2.2.2

CHAPTER 2: METHODOLOGY

Our project can be divided into two parts. The first part is the building classic Chinese English Dictionary. The second part is evaluating and analyzing the modern Chinese Hungarian dictionary built by joining of a published Chinese-English dictionary and an English-Hungarian dictionary already built by the SZTAKI Dictionary project [5].

2.1 Classic Chinese-English Dictionary

2.1.1 Collecting Parallel Corpus

The Chinese Text Project [3] maintained a website that recorded large amount of text resources related to classic Chinese such as the texts from Confucianism, Mohism, Daoism, Legalism and other classic Chinese philosophy. Part of the text resources were translated to English in a crowdsourcing manner. The tool that we used to generate classic Chinese to English dictionary requires parallel corpus of those two languages. Therefore, we utilized this website and downloaded the necessary content.

The Chinese Text Project provided different web page formats to display the classic Chinese texts and their corresponding English translations. For the first format, the texts were grouped by paragraphs, and for the second format, the texts were grouped by sentence. The text grouping is important because the tool that we used to generate the dictionary, Hundict [30], required the input Chinese to English parallel corpus to be aligned by sentences. So we decided to download the web pages in the second format. Also, for the second format, each Chinese word is marked with a hyperlink to Chinese Text Project's built-in dictionary. In this way, each Chinese word is separated by the hyperlink tag in the web pages' HTML file. This is also an

important characteristic because Hundict distinguishes different words by spaces. Commonly Chinese texts are concatenated without spaces. With each Chinese word being separated by a hyperlink tag in the HTML file, when parsing the HTML file, we can easily split each word and put spaces between them.

There are automated tools, such as wget (https://www.gnu.org/software/wget/) [27], to help people download website content conveniently. The Chinese Text Project contained thousands of valuable web pages for dictionary generation. Thus, using such download tools would greatly save our time and energy. However, The Chinese Text Project had an efficient protection mechanism to prevent users from downloading its website content by such automated downloading tools. Therefore, we manually downloaded the content we needed using a common browser, instead of utilizing such tools.

We downloaded 2636 web pages that contain Chinese and English alignment, and obtained 342249 pairs of classic Chinese to English sentence translations.

In addition, despite that the Chinese Text Project prevents people from downloading its content by automated download tools; automated download is still achievable by other methods. Normally, the Chinese Text Project browser does not reject access requests from a common browser; therefore, one can build a mouse macro to automate the download action. We had attempted this approach.

In order to download the web pages with the second format, we need to go to the web page with the first format that groups the class Chinese to English translations by paragraphs. At the beginning of each paragraph, there were a set of function buttons, and one of the buttons was marked with ‘jump to dictionary’ tag. By clicking such buttons, the web pages will be navigated

to the web pages with the second format. Also, one can right click such buttons to perform the download action to download the proper web pages in HTML or HTM file extensions.

OpenCV (http://opencv.org/) [19] (Open Source Computer Vision Library) is an open source computer vision and machine learning software library. Using OpenCV, one can easily build tools detecting the position of an object within an image. We utilized OpenCV library and built an automated download tool for the Chinese Text Project's website.

When the Chinese Text Project's web page with the first format was open, we can see several navigation buttons, each located at the beginning of the paragraph. If we were then to manually download the web pages with the second format, we need to right click on those buttons and hit the save button on the pop-up menu. Therefore, to construct the automated download tool, we first used python to take a snapshot of the screen, and used OpenCV library to detect the positions of every buttons. With this position information, we can use the mouse macro to automatically move the mouse to each of the positions where the buttons were located on the screenshot, then perform the right click, again, using the mouse macro. After the right click, we can repeat taking the screenshot and let OpenCV to detect the save button's position in the pop-up menu, and move the mouse to the position of that save button and perform a left click to finish the download of one page. Repeating these steps will download all the content we are interested in within the current web page. To achieve the mouse macro, we built a Java server that received the position information of where the mouse should be moved to by User Datagram Protocol (UDP) [21], and which mouse button should be clicked, and perform those actions.

The automated download tool was not complete and useful. It can only download the contents that are linked by the navigation buttons on the current page. In order to complete the

download tool, the tool needs to be able to automatically detect the location of the link for the next chapter of the classic Chinese texts. While detecting navigation buttons can be simply achieved by calling the built-in method of OpenCV library, recognizing the hyperlinks on a screenshot would require more knowledge about OCR [17], which is out of the scope of automated dictionary generation.

2.1.2 Extracting data from webpage

The web pages that we downloaded were in HTML or HTM format, written by markup languages. By examining the source code of such files, we found out the pattern that the data that we were interested in was always within a table tag that carries the ‘id’ as attribute and has the attribute value ‘translationtable’. Within the table tag, there were classic Chinese texts and corresponding English texts; each of them was within a separate table tag. Each Chinese word was inside of a separate ‘a’ tag, which was a hyperlink towards the dictionary provided by the website. Generally Chinese words are written without a space within a sentence, but in this way, the Chinese texts from the Chinese Text Project web page already had words segmented, which was an important characteristic as the input of the following procedures. We built python tools, mainly utilizing the python library HTMLParser, to detect and extract texts using this pattern, and produced a text file with 34140 lines. Within each line, there was a classic Chinese sentence on the left and an English sentence on the right. A tab character separated the two sentences.

The parsing tool that we built cannot completely parse the entire HTML file that we downloaded. There were rare cases that the parsing tool could not recognize table tags in the HTML file and cases that the Chinese Text Project’s Chinese to English alignments were inaccurate or incorrect. Therefore the output of the parsing tool could have certain defects on

alignment. For example, some of the lines were missing the tab characters between the Chinese and English. To remove those defect lines, we built a script to determine which lines contain both Chinese Characters and English characters. The characters that were within the coding range from u'\u4e00' to u'\u9fff' should be determined as Chinese characters [28]. After filtering out the defect lines, we manually corrected those lines. The reason we did not use an automated script to fix those incorrect lines was that there were other unpredictable alignment error types.

2.1.3 Word-Level segmentation

Commonly, words in Chinese texts are not separated by spaces as English texts do. To preprocess such Chinese texts to split the Chinese words with spaces, we utilized the Stanford Word Segmenter (http://nlp.stanford.edu/software/segmenter.shtml) [25]. However, the Stanford Word Segmenter is trained on modern Chinese corpus [12]; therefore the result of the Stanford Word Segmenter is not accurate when processing classic Chinese texts.

Because the Chinese Text Project provided us a web page format with each Chinese word segmented by the hyperlink tag in the HTML file, the Stanford Word Segmenter is not necessary for preprocessing the corpus downloaded from the Chinese Text Project.

2.1.4 Word-Word Classic Chinese dictionary using Hundict

Hundict is a tool that takes a list of parallel sentences in two languages and produces a dictionary based on the frequency that a particular word occurs with other words in the same line of the parallel sentences. Generally, if a Chinese word always comes with another English word in the corresponding English sentence, these two words will have a high chance to have the same meaning. After applying Hundict to the parallel corpus we obtained, it produced a dictionary

with 1064 entries, marked with confidence values. These dictionary entries needed further evaluation.

2.2 Evaluate and Analyze Modern Chinese-Hungarian Dictionary

In this section, we talk about how we extracted the Chinese-English dictionary from a MDB file and joined that with an English-Hungarian dictionary to get a Chinese-Hungarian dictionary. Then we discuss the details of the methodology and techniques we used to evaluate and improve our bilingual dictionary between modern simplified Chinese and Hungarian.

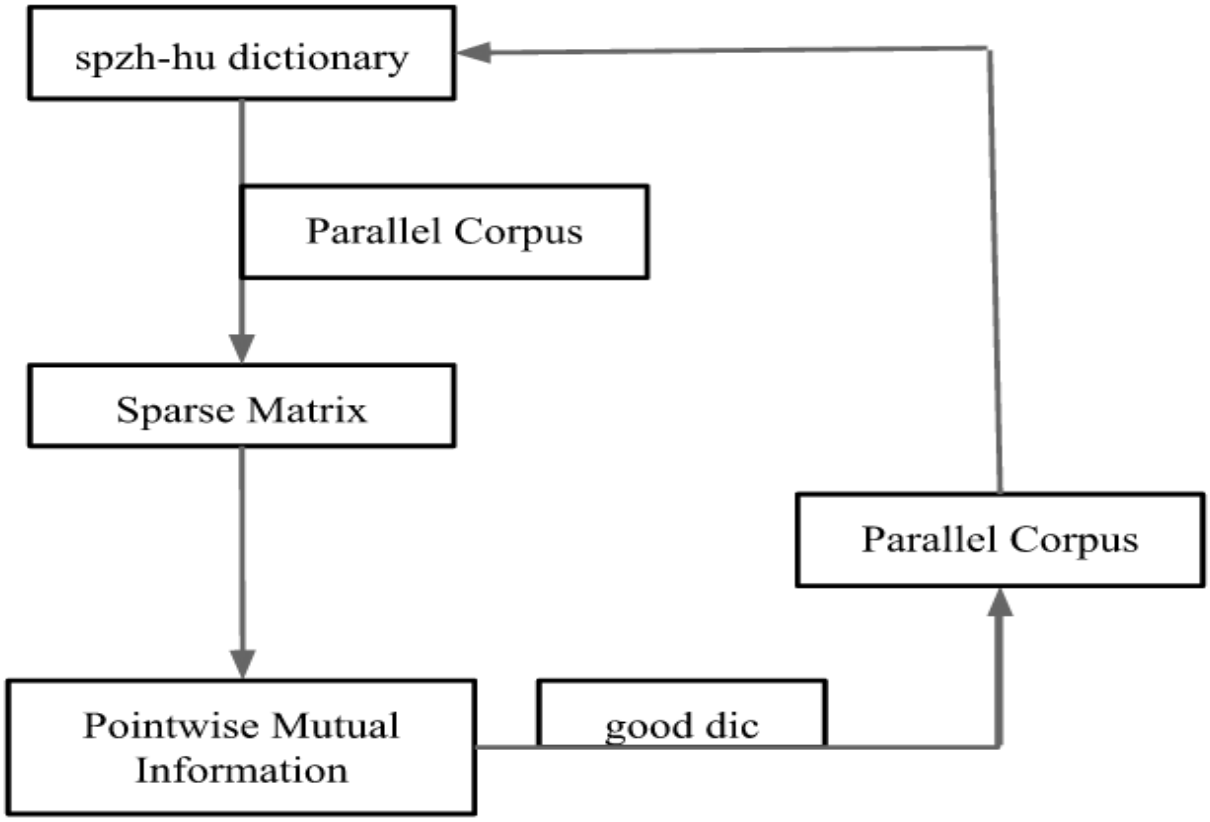

Figure 2: Methodology Flowchart of spzh-hu dictionary

Figure 2 above summarizes the methodology used to evaluate and improve our bilingual dictionary between modern Chinese (spzh) and Hungarian (hu). In general, we joined a

published English-Chinese dictionary with an English-Hungarian dictionary (previously done by 4lang group) [1] to get a Chinese-Hungarian dictionary. Then we analyzed the dictionary by building a sparse matrix and analyzing the Pointwise mutual information.

2.2.1 Building a Dictionary by joining two dictionaries

The modern Chinese – Hungarian dictionary we used in this project was built by joining a published Chinese-English dictionary and an English-Hungarian dictionary built by the SZTAKI Dictionary project [5].

After obtaining a published Chinese-Hungarian dictionary, which was a MDB file, we used some extraction software such as *MDB Explore* and *AccessLook* to read and to extract the data from MDB files. That software allowed us to extract MDB data to some simpler format, for example: txt, sql, xml. After that, we wrote a python program to clean the data to a file zh-en_dic.txt with the format we could easily use: *English word* + ' @ '+ *Chinese Word*.

To join the Chinese-English dictionary and English-Hungarian dictionary, we used the *join* command in linux (man page: http://linux.die.net/man/1/join). To use the *join* command, the two files need to be in a certain order. For each dictionary, we used the *sort* (man page: http://linux.die.net/man/1/sort) command to sort the pairs in alphabetical order according to the English word. By joining two dictionaries, we got a dictionary with 2,109,727 candidate pairs.

2.2.2 Applying a Sparse Matrix

We used a sparse matrix to evaluate our dictionary. For each candidate pair (*zh1-hu1*) in the dictionary, we traversed the whole aligned parallel corpus [29] sentence by sentence. In each sentence, we recorded how many times both the Chinese word (*zh1*) and the Hungarian (*hu1*)

appeared; how many times the Chinese word (*zh1*) appeared but the Hungarian word (*hu1*) did not; how many times the Hungarian word (*hu1*) appeared but the Chinese word (*zh1*) did not. Therefore, our sparse matrix looks like the following:

	candidate pair 1 *zh1-hu1*			candidate pair 2 *zh1-hu2*			candidate pair 3 *zh2-hu3*			…
	both	only *zh1*	only *hu1*	both	only *zh1*	only *hu2*	both	only *zh2*	only *hu3*	…
s1	3	1	0	2	0	0	1	0	3	…
s2	1	0	2	3	11	0	2	3	0	…
…										…

Table 1: Sample of our sparse matrix

2.2.3 Pointwise Mutual Information

After we got the sparse matrix, we summed each column to get the total times each word appears in the whole parallel corpus. For each candidate pair in the dictionary, we were able to build a matrix [see Table 2].

	zh	¬ *zh*
hu	n_b	n_h
¬*hu*	n_z	

Table 2: Template matrix for each phrase in the dictionary

n_b stands for the number of times both *hu* and *zh* appear;

n_h stands for the number of times only *hu* appears;

n_z stands for the number of times only *zh* appears

Through this matrix, we were able to calculate the PMI of the pair:

$$PMI(zh, hu) = \log\frac{p(zh, hu)}{p(zh)p(hu)} = \log\left(\frac{\frac{n_b}{N}}{\frac{n_z + n_b}{N} * \frac{n_h + n_z}{N}}\right)$$

$$= \log\left(\frac{n_b}{(n_z + n_b) * (n_h + n_b)} * N\right)$$

Here N is the total number of tokens that appeared in the whole parallel corpus, which was 8,582,791(number of Hungarian words) + 1,782,380 (number of Chinese characters), $n_z + n_b$is the total number of appearances of the Chinese word; $n_h + n_b$ is the total number of appearances of the Hungarian word. Thus, for the convenience of calculating, both <zh> and <hu> in our program were the total number of appearances of either Chinese or Hungarian. Figure 3 is the first thirteen lines of the sum of our sparse matrix. Taking our first pair into example, the matrix is [see Table 3].

```
African @ 非洲人 @ afrikai :  <pair> 3 <zh> 6 <hu> 45
American @ 美国人 @ amerikai :  <pair> 117 <zh> 220 <hu> 455
American @ 美洲人 @ amerikai :  <pair> 1 <zh> 6 <hu> 455
Anglo-Saxon @ 盎格鲁撒克逊 @ angolszász :  <pair> 1 <zh> 2 <hu> 14
April @ 四月 @ április :  <pair> 70 <zh> 104 <hu> 173
April @ 四月份 @ április :  <pair> 8 <zh> 11 <hu> 173
Arab @ 阿拉伯 @ arab :  <pair> 33 <zh> 70 <hu> 2639
Arab @ 阿拉伯人 @ arab :  <pair> 18 <zh> 31 <hu> 2639
Arab @ 阿拉伯马 @ arab :  <pair> 1 <zh> 1 <hu> 2639
Arabian @ 阿拉伯 @ arab :  <pair> 33 <zh> 70 <hu> 2639
Arabian @ 阿拉伯人 @ arab :  <pair> 18 <zh> 31 <hu> 2639
Asian @ 亚 @ ázsiai :  <pair> 10 <zh> 8875 <hu> 18
Asian @ 亚洲 @ ázsiai :  <pair> 8 <zh> 45 <hu> 18
```

Figure 3: First thirteen lines of the sum of spars matrix

	非洲人	¬ 非洲人
afrikai	3	45-3
¬afrikai	6-3	

Table 3: An example applying PMI

$$PMI(\text{非洲人}, \text{afrikai}) = \log(\frac{3}{6 * 45} * 10{,}365{,}171) = 14.0012952777$$

The PMI of the candidate pair (非洲人, afrikai) is 14, which is considered to be very high. Checking with someone who speaks Hungarian, the word ‘afrikai’ means ‘African’; meanwhile, ‘非洲人’ also means ‘African’. As a result, this candidate pair could be regarded as a correct pair. Already having PMI for all twenty thousand candidate pairs, we needed to find out the boundary between good candidate pairs and bad pairs. After sorting the pairs in increasing order of PMI and checking with professor Huba Bartos, who is a Chinese expert Hungarian, we found out that the pairs with a PMI around or greater than 6, with total number around fifteen thousand, had very high precision. There were several kinds of ‘bad’ pairs. We will talk about this in Section 3.2

CHAPTER 3: RESULTS AND EVALUATION

3.1 Result 1: classic Chinese-English dictionary

The resulting classic Chinese to English dictionary contained 1196 entries. Each entry contained a confidence value, a Chinese word and a corresponding English word.

By manually labeling the incorrect entries, we obtained the precision of the generated classic Chinese to English dictionary. There were 508 entries with confidence value of 3.0 and above, and the precision for them is 96%. For the 598 entries with confidence value of 2.5, the precision is 94%. For the 678 entries with confidence value of 2.0, the precision is 93%.

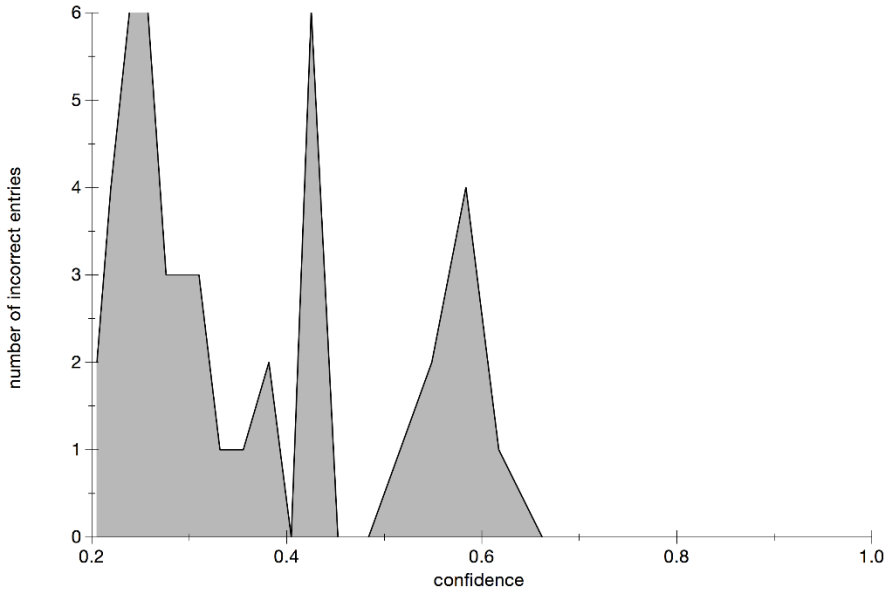

Figure 4: Distribution of Error in clzh-en dictionary with confidence above 0.2

The figure above showed the frequency of incorrect translation in our generated dictionary against the confidence value. According to the figure, the incorrect translation mainly distributed at low confidence value interval. We cut off the translations that with a confidence

value below 0.2, 518 translation entries in total, because the quality of those translations were not good enough, and also because the labor work of manually verifying the correctness of translations with low confidence value was more intense than those with high confidence value.

The first incorrect translation in the dictionary file, after the dictionary's translation entries were sorted by the confidence value, was 成汤 @ complet. As the first translation we spotted to be incorrect, this translation had a confidence value of 0.61. First of all, the word complet is derived by stemming the word completer. The Chinese word 成汤 is a name. By examining the input corpus for Hundict, we found out that in the corpus from the Chinese Text Project, the name 成汤 was translated to Tang the completer, therefore the frequency of 成汤 and Tang pair is equal to the frequency of 成汤 and completer pair, therefore Hundict produced incorrect result. Similarly, the translation of 代 @ Three was incorrectly generated by the same reason. The Chinese character 代 in the corpus was separated from the original word 三代圣王, literally means the Great kings or rulers from the three dynasties, referring to the ancient rulers named Yao Shun Yu Tang Wen and Wu. The translation given by the corpus was Three Dynasties; therefore the frequency of the word three and dynasty were the same, leading the incorrect translation to 代. The confidence value for this translation pair was 0.59. However, after running Hundict with its ngram option, the resulting dictionary contained the translation of 三代 @ Three Dynasties, which was perfectly correct, with a high confidence of 0.97.

The second translation error with high confidence value was 诸@ feudal, which should be 诸侯 @ feudal lord. The Hundict sometimes failed to recognize the complete English or Chinese phrase, so that it produced misleading translation.

The third translation error was caused by the misleading of related words. For example, in a Chinese phrase: rise up at dawn and retire in the night, retire and dawn were related with respect to the frequency, so that Hundict produced translation from Chinese word ‘retire’ to English word ‘dawn’.

The incorrect translations that with the confidence value from 0.55 to 0.65, which was the first pike in the figure above, were mostly caused by the reason discussed above.

In the figure above, the second narrow pike of errors were mainly caused by the third translation error.

For evaluating the recall for our automated generated dictionary, we used two existing dictionaries. One dictionary is called ‘basic dictionary’, which only contained 100 basic words such as cat, women, and freedom. Those words were important but simple. They could be used to define other more complex words. The second existing dictionary was English to modern Chinese dictionary from the 4lang project [1].

We used the ‘join’ command to select the dictionary entries that present both in the basic dictionary and our generated dictionary, and then obtained a recall of 42%. Also, by joining our generated dictionary with English to Chinese dictionary from the 4lang project, we obtained a recall of 22%. The main reason for the low recall is that the English stemming tool produced incomplete English words in our dictionary, such as ‘complet’, ‘Princ’, ‘Someon’. The joining method cannot recognize these incomplete words, therefore reduced the calculated recall. Therefore, our attempt to evaluate the recall for the dictionary should be considered incomplete and future work on fixing the incomplete words is necessary.

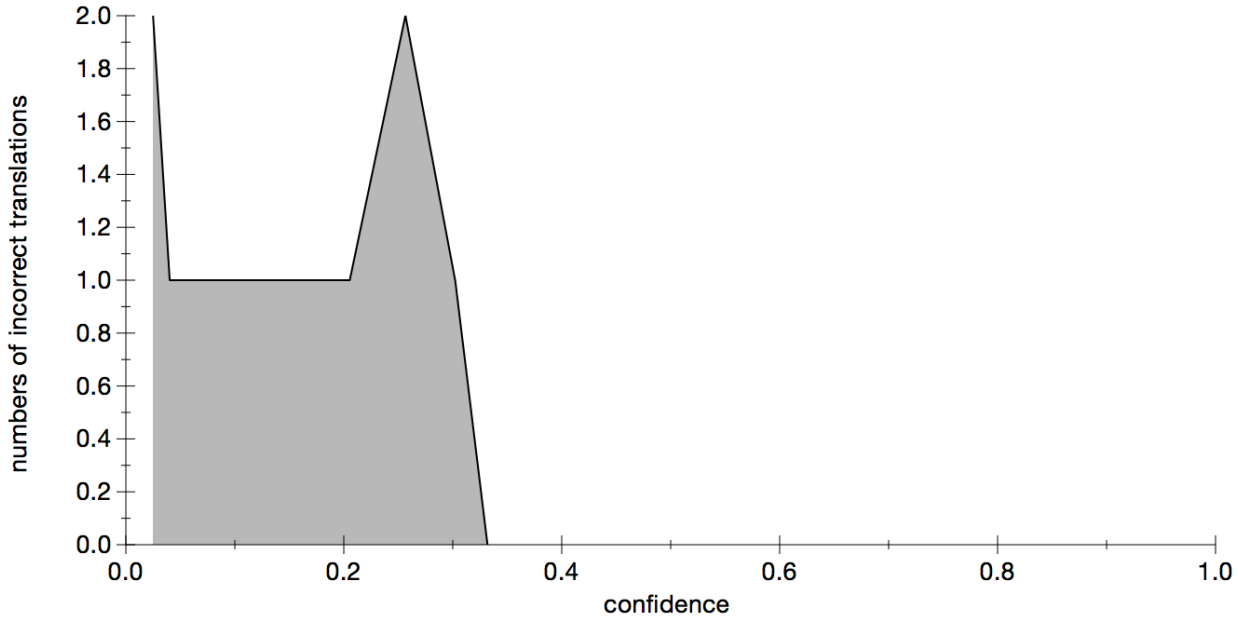

Figure 5: Distribution of Error in clzh-en dictionary

For the result of joining our classic Chinese to English dictionary with the 100 words basic dictionary, the precision of the translation with confidence above 0.2 is 94%. As showed in the figure above, most of the incorrect translation occurred in the section with low confidence, and the high confidence translations remained good quality.

3.2 Result 2: Modern Chinese – Hungarian Dictionary

3.2.1 Description of the Chinese-Hungarian Dictionary

The final Chinese-Hungarian Dictionary was a list of candidate pairs filtered by PMI greater than or equal to 6. Each pair contained two parts. In the first part, it stored the different expression of one single meaning, with the order of English, Chinese and Hungarian. The second part is the result of the sparse matrix and PMI. Some good examples are below:

```
ailment @ 病 @ betegség :  <pair> 318 <zh> 2868 <hu> 472 <PMI> 8.457558971?1
aim @ 对准 @ cél :  <pair> 18 <zh> 148 <hu> 2411  <PMI> 6.2182040742
aim @ 目标 @ cél :  <pair> 108 <zh> 369 <hu> 2411 <PMI> 7.48514293449
aim @ 目的 @ cél :  <pair> 257 <zh> 1038 <hu> 2411        <PMI> 7.24376625922
aim @ 目的地 @ cél :  <pair> 48 <zh> 173 <hu> 2411 <PMI> 7.40806671147
aim @ 瞄准 @ cél :  <pair> 66 <zh> 175 <hu> 2411  <PMI> 7.85091544591
aim @ 针对 @ cél :  <pair> 4 <zh> 63 <hu> 2411    <PMI> 5.28045251488
air @ 神态 @ arckifejezés :  <pair> 6 <zh> 309 <hu> 270   <PMI> 6.72983224975
air @ 神气 @ arckifejezés :  <pair> 16 <zh> 573 <hu> 270  <PMI> 7.25394144818
airfield @ 机场 @ repülőtér :  <pair> 30 <zh> 51 <hu> 33  <PMI> 14.6832195083
airplane @ 飞机 @ repülőgép :  <pair> 24 <zh> 193 <hu> 35 <PMI> 12.3563708205
aisle @ 走廊 @ folyosó :  <pair> 386 <zh> 738 <hu> 1124   <PMI> 9.42370208497
```

Figure 6: Example of candidate pairs in the final dictionary

The whole dictionary was sorted in alphabetical order by the English expression of the word. If an English word had different meanings, we considered those to be separate pairs

3.2.2 Distribution of the result:

The total distribution of the whole original Chinese - Hungarian Dictionary has more than twenty thousand candidate pairs. The distribution of PMI [see Table 4] is below. We randomly selected twenty sample pairs from different PMI intervals, checked the precision manually.

From our manual check, the overall precision was increasing when the PMI increased. The precision of sample pairs in interval [2, 4) was 20% (4 out of 20) and the precision of sample

pairs in interval [6, 8) was 85% (17 out of 20). The precision of sample pairs with PMI greater than 8 were almost all correct. As a result, we considered PMI greater than 6 as good pair, with the total amount of 14573 pairs.

PMI Range	Number of Pairs	Percentage
< 0	5	0.00024521
[0, 2)	532	0.02608994
[2, 4)	3403	0.16688735
[4, 6)	3536	0.17340984
[6, 8)	3892	0.19086852
[8, 10)	3772	0.18498357
[10, 12)	2504	0.12279927
[12, 14)	1467	0.07194350
[14, 16)	802	0.03933108
[16, 18)	374	0.01834143
[18,20)	83	0.00407042

Table 4: PMI distribution

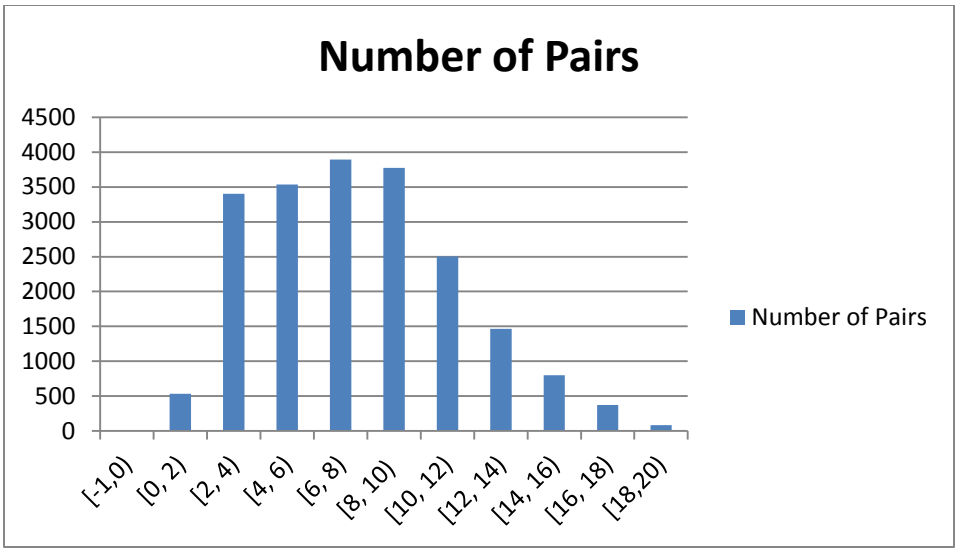

Figure 7: Column Chart of PMI Distribution

3.2.3 Common Errors

Here are some common errors we found in the low PMI area:

Error I: Wrong pair. This is the main type of bad candidate pairs. The Chinese word and the Hungarian word had totally different meanings. Some examples were: 'gift @ 天分 @ ad', 'silent @ 密 @ csendes. 'ad' means *give*, while '天分', means *gift* or talent; 'csendes' means *quiet*, but '密' means *secret* or *dense*. The main reason for these errors was that either our Chinese-English dictionary or our English-Hungarian dictionary or both dictionaries had wrong entries.

Error II: Slang expression. Some words, when used orally, have a totally different meaning from their primary meaning. Some examples are: 'stupid @ 二 @ buta'. The word '二' in oral Chinese has a meaning of 'stupid' when joking with a friend. In written Chinese there is no such usage. Another example is: 'bird @ 女人 @ madár'. Both the English word 'bird' and

the Hungarian word ‘madár’ primarily mean the vertebrate with two wings. People say ‘a pretty bird’ to praise a beautiful woman. However, ‘bird’ does not mean ‘woman’. Although candidate pairs in this kind of error were not technically wrong, these pairs were not formal and should be removed.

Error III: Inconsistent in part of speech. A good example is in figure 6. All these Chinese words (11) have the meaning of *anger* and the Hungarian word ‘bosszant’ also means *anger*. However, ‘bosszant’ is a verb while only 6 out of 11 of Chinese words (‘发怒’, ‘愤’, ‘怒’, ‘触怒’, ‘激怒’ and ‘生气’) are verbs. 3 out of 11 (忿, 怒气, 怒火) are nouns; and the remaining two (气愤, 愤怒) are adjectives.

```
anger @ 发怒 @ bosszant : <pair> 1 <zh> 71 <hu> 302 <PMI> 6.10499651516
anger @ 忿 @ bosszant : <pair> 1 <zh> 147 <hu> 302 <PMI> 5.05507128982
anger @ 怒 @ bosszant : <pair> 27 <zh> 2766 <hu> 302 <PMI> 5.57604569567
anger @ 怒气 @ bosszant : <pair> 1 <zh> 307 <hu> 302 <PMI> 3.99264878929
anger @ 怒火 @ bosszant : <pair> 1 <zh> 265 <hu> 302 <PMI> 4.20489508521
anger @ 愤 @ bosszant : <pair> 5 <zh> 1461 <hu> 302 <PMI> 4.06393126674
anger @ 愤怒 @ bosszant : <pair> 3 <zh> 769 <hu> 302 <PMI> 4.25286634742
anger @ 气愤 @ bosszant : <pair> 2 <zh> 289 <hu> 302 <PMI> 5.07981795216
anger @ 激怒 @ bosszant : <pair> 9 <zh> 152 <hu> 302 <PMI> 8.17674112266
anger @ 生气 @ bosszant : <pair> 23 <zh> 1002 <hu> 302 <PMI> 6.80963879752
anger @ 触怒 @ bosszant : <pair> 1 <zh> 23 <hu> 302 <PMI> 7.7311816786
```

Figure 8: Example of Inconsistent in Part of Speech Error

Error IV: Indirect translation. In our dictionary, there was a pair ‘life @ 春 @ energia’. The direct meaning of Chinese word ‘春’ is spring. But spring is full of vitality. Chinese always use ‘春’ to represent ‘life’. However, ‘life’ could not be replaced by ‘spring’ in application: there is no usage such as ‘spring is hard’ to represent the meaning of ‘life is hard’.

CHAPTER 4: CONCLUSION

In this project, we created an automatic machine generated Classic Chinese – English Dictionary and a Chinese – Hungarian dictionary. We presented a methodology for extracting parallel corpora and dictionaries between character-based languages and word-based languages.

Our project is an important contribution and improvement to the well-reputed SZTAKI Szótár (‘Dictionary’ in Hungarian) project (Sztaki Dictionary) [5]. Currently, SZTAKI Szótár collected bilingual dictionaries in 7 languages: English, Hungarian, German, French, Italian, Polish, Dutch and Bulgarian. However, all the 7 languages are European word-based languages. The introduction of Chinese, a language spoken by about one fifth of the world’s population, would significantly increase the diversity and influence of SZTAKI dictionaries. Furthermore, a great number of scholars are studying works in Classic Chinese. Implementing Classic Chinese dictionary would be an important resource for researchers for understanding ancient Chinese.

CHAPTER 5: FUTURE WORK

For the classic Chinese to English dictionary, since the English stemming tool produces incomplete English words, using a better stemming tool could improve the quality of the dictionary and fix the error when calculating recall.

Commonly the Chinese corpus does not use space to separate each word. Therefore word segmentation is often necessary before one can apply tools like Hundict on such corpus. As most Chinese segmentation tools available are trained with modern Chinese corpus, generally the quality of segmenting classic Chinese corpus is not ideal. Since the corpus downloaded from the Chinese Text Project already has the word segmentation, it is possible to utilize such this segmentation data to implement a classic Chinese segmentation tool.

There is another possible improvement for the classic Chinese dictionary. In some Chinese phrases or expressions, each character in that phrases or expressions are equal, so that each character has the same frequency associate with the corresponding English translation. Therefore, there are errors that Hundict aligned incorrect character in the phrases or expressions with the English word. This type of error can be further studied can be eliminated by filter out the translations from those phrase or expression from the dictionary to improve the precision.

For modern Chinese dictionary, it is necessary to improving and extending the dictionary using more parallel corpus. Moreover, future research should work on building a traditional Chinese – Hungarian dictionary which is widely used in Taiwan, Hong Kong.

Using the same methodology, we will be able to building more bilingual dictionaries involving other character based language such as Japanese – Hungarian dictionary or Hungarian – Korean dictionary.

Bibliography

[1] Andra´s Kornai and Ma´rton Makrai. ‘A 4lang fogalmi szo´ta´r [the 4lang concept dictionary].’ Tana´cs and V. Vincze. Magyar Sza´mito´ge´pes Nyelve´szeti Konferencia [Ninth Con- ference on Hungarian Computational Linguistics]. 2013. 62-70.

[2] Attila Balogh, Zsolt Both, András Farkas, Péter Halácsy. http://www.hunglish.hu/.

[3] Chinese Text Project, http://ctext.org/

[4] Church, Kenneth Ward, and Mark Y. Liberman. 1991. A status report on the ACL/DCI . in Proceedings of the 7th Annual Conference of the UW Centre for New OED and Text Research: Using Corpora. Computational Linguistics 19:1-24

[5] Computer and Automation Research Institute, Hungarian Academy of Science. Sztaki Dictionary. Budapest

[6] Fano, Robert M. 1961. Transmission of information; a statistical theory of communications. New York: MIT Press

[7] Golub, Gene H.; Van Loan, Charles F. Matrix Computations (3rd Ed.). Baltimore: Johns Hopkins, 1996.

[8] Grigg, Hugh. Past events in Mandarin Chinese grammar. 14 April 2013. 22 4 2015.

[9] Grime, B. The Ethnologue (14th Edition).SIL. 2003.

[10] Gut, Allan (2013). Probability: A Graduate Course (2 ed.). New York, NY: Springer.

[11] Hindle, Donald. 1990. Noun classification from predicate argument structures. In ACL 28, pp.286-275

[12] Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, Christopher Manning. ‘A Conditional Random Field Word Segmenter, for Sighan Bakeoff 2005.’

[13] Jeff Bolz, Ian Farmer, Eitan Grinspun, Peter Schrooder. ‘Sparse matrix solvers on the GPU: conjugate gradients and multigrid.’ ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2003 (2003): 917-924.

[14] Kay, M., Roscheisn, M. . Technical Report P90-00143, Xerox Palo Alto Research Center. . 1988. < http://acl.ldc.upenn.edu/J/J93/J93-1006.pdf>.

[15] Li Haizhou, Yuan Baosheng. Chinese Word Segmentation. Singapore: Language, Information and Computation, 1998.

[16] Liang Tian, Derek F. Wong, Lidia S.Chao, Paulo Quaresma, Franciso Oliveira, Yi Lu, Shuo Li, Yiming Wang, Longyue Wang. ‘UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation.’ 2014.

[17] Mori, S.;Suen, C.Y. ; Yamamoto, K. ‘Historical review of OCR research and development.’ 06 Augest 2002.

[18] Nie, J., Simard, M., Isabelle, P., Durand, R. cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In Proceedings of the 22nd Annual International ACM-SIGIR. Web. 1999.

[19] OpenCV http://opencv.org/.

[20] Parallel Text, http://en.wikipedia.org/wiki/Parallel_text.

[21] Postel, J. User Datagram Protocol. 1980.

[22] Potthast, M., Barron-Cedeno, A., Stein B., Rosso, P. Cross-Language Plagiarism Detection. 2010.

[23] Resnik, P., Smith, N. The Web as a Parallel Corpus. Computational Linguistics. Sep, 2003.

[24] Salkie, Raphael. Using parallel corpora in translation. 14 4 2015.

[25] Stanford Word Segmenter http://nlp.stanford.edu/software/segmenter.shtml. 20April 2015.

[26] Timothy A. Davis, Yifan Hu. 'The university of Florida sparse matrix collection.' ACM Transactions on Mathematical Software (2011).

[27] wget https://www.gnu.org/software/wget/.

[28] Yang, Herong. U4E00: CJK Unified Ideographs. 2012. 22 4 2015.

[29] Zhongxiu Liu, Yidi Zhang. 'Automated Building of Sentence-Level Parallel Corpus and Chinese-Hungarian Dictionary.' 2013.

[30] Zséder, Attila. https://github.com/zseder/hundict