澳 門 理 工 學 院
Instituto Politécnico de Macau
Macao Polytechnic Institute

School of Public Administration
Bachelor of Science in Computing

COMP491 Final Year Project
Final Report

Academic Year 2016/17

Bilingual Corpus Construction System based on Innovative Length-based Sentence Alignment Algorithm

Project number:	2
Student ID:	P-13-0836-9
Student Name:	Xu Jiabo Bob
Supervisor:	Dr. Sio Tai Cheong, Victor
Assessor:	Dr. Rita Tse
Submission Date:	April 19, 2016

Declaration of Originality

I, Xu Jiabo, declare that this report and the work reported herein was composed by and originated entirely from me. This report has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given in the bibliography.

许家博

2017.3.15

Abstract

Machine Translation is a very popular topic in Natural Language Processing in the last few decades. Since the twentieth century, many organizations started constructing multilingual corpus, such as Europarl, ASPEC (Asian Scientific Paper Excerpt Corpus), News Commentary corpus etc. Translations between most European languages such as English, Greek, Spanish, are ubiquitous on internet. Moreover, Chinese-English corpora, such as Corpus Linguistics at BFSU (Beijing Foreign Study University), have also been flourished years before. However, there is no considerable scale of bilingual corpora concerning Chinese-Portuguese.

This project covers the five major processes of Chinese bilingual corpus construction, including data collection, data cleaning, sentence alignment, translation evaluation and cloud platform. It accomplished a bilingual corpus construction process from bilingual corpus collection to human modification platform.

In the data collection section, web crawler is the essential tool that is used for gathering bilingual data from candidate websites. Three types of website structures and their corresponding crawl techniques will be introduced in this report. With respect to data cleaning, this project mainly deals with the special cases on word and sentence split, which frequently occur in most European languages. Moreover, HTML tag cleaning and some other noise removing processes are included. Furthermore, a length-based sentence alignment algorithm is designed and verified. Compared with Gale and Church's algorithm, this algorithm can break up Chinese sentence into reasonable small size. Finally, it got 84.5% passing rate out of 105 TED speeches. The passing rate is generated from a full-automatic translation evaluation metric which is also firstly attempted in this project. It regards the combination of Baidu and Google translation results as the reference of Bilingual Evaluation Understudy (BLEU) score. Last but not least, integrating with all these above tools, this project provides a cloud platform for human checking and modifying the corpus.

To sum up, this project can be considered as an innovative and reasonable attempt in setting up the bilingual corpora with Chinese and Portuguese, and it has solved the problems at the initial stage of the corpus construction.

Acknowledgement

This report would not exist if not for the support and guidance of a special group of people who will always have my deepest appreciation.

I would like to thank Dr. Sio Tai Cheong, Victor, my project supervisor, who has offered his/her technical supports and guided me into this research.

I would like to thank Dr. Rita Tse who has been the assessor of my report.

I would like to thank my classmates and friends in the Computing Program of Macao Polytechnic Institute.

I cannot find words strong enough to express my appreciations for my family.

Table of Contents

Table of Figures

List of Tables

1 Introduction

Corpus Linguistics, which is the study of language as expressed in corpora (samples) of "real world" text [1], has revived and flourished since the eighties of twentieth century [2]. Corpora are the fundamental material which is utilized to find invisible linguistic relationship behind. Thus, corpora are the priceless treasure for people whose study is related to linguistics. So far, there have been many famous corpora around the world. In terms of English corpora, British National Corpus (BNC) released by British, BROWN corpus constructed by Brown University and Linguistic Data Consortium(LDC) established by University of Pennsylvania are the most influential large scale corpora in the world. Considering Chinese, due to the promotion of Chinese international status, Chinese becomes one of the frequently used languages in the world. Meanwhile, Chinese text information becomes much larger than before. The most famous corpus level collection coming from Peking University is Center for Chinese Linguistic which includes modern Chinese, ancient Chinese and English.

Parallel Corpus, which contains different translations for each single document, is one common type of corpus. Machine Translation is one of the edging research fields that particularly demands bilingual corpus. Together with the prosperity of Corpus Linguistics, Machine Translation has become a very popular topic in Natural Language Processing during the past few decades. Since the twentieth century, many organizations have started constructing multilingual corpus, such as Europarl [3], ASPEC, News Commentary corpus etc. Translations between most European languages and English are ubiquitous on internet. Moreover, Chinese-English corpora, such as Corpus Linguistics at BFSU, have also been flourished years before. However, there is no considerable scale of bilingual corpora concerning Chinese-Portuguese, since bilingual corpus establishment comprises complex procedures. Besides, the computer only plays an auxiliary role, while manual processing will be sharply decreased.

Parallel corpus construction, therefore, becomes a very basic task, particularly in the initial stage. Whereas, more complex than monolingual corpus, parallel corpus should satisfy the alignment at least in the sentence level, which means each sentence in one language should have its corresponding sentence in another language. Comparatively, Corpus which is aligned in document level is less valuable. Therefore, the construction procedure, including three central steps, aims at establishing corpus with parallel sentence. Those steps are data collection, data cleaning and sentence alignment. In addition, the final result should be not only easy to read, but also liable to convert into an available form of translation model.

In order to construct parallel corpus, sentence alignment is an inevitable technique. The statement from Wiki declares that “In parallel corpora single sentences in one language can be found translated into several sentences in the other and vice versa.” [4] There are many researches on sentence alignment for European languages. However, different from English-like languages, such as French and German, Chinese has extremely different syntax and punctuation system. Besides, the concept of sentence in Chinese is ambiguous. Thus, some methods do not perform well on Chinese. As a result, in order to construct a Chinese bilingual corpus, a dedicated sentence alignment method is demanded.

1.1 Objectives

The goal of this project is to build a reliable environment for constructing parallel corpus. It contains five procedures, which are data collection, data cleaning, sentence alignment, evaluation metrics and modification platform.

The objectives and corresponding outcomes are indicated as follows:

1.1.1 Collecting monolingual or bilingual corpus from the Internet

In the very first stage, it is urgently demanded that the channel, which is used to obtain raw corpus, must be efficient and reliable. Thus, traditional methods, such as artificial collecting, cannot be the sustainable solution. Web crawler, to some extent, is the best alternatives. The web crawler should be durable enough so that it can supply new corpus in long a period of time. Besides, in order to collect full-scale corpus, more than one website is needed to be crawled. Therefore, the crawler should have the ability to crawl from different website at the same time. Furthermore, the raw result must at least be document aligned, otherwise it will be more consumptive if align them in later procedure.

1.1.2 Fulfilling robust data pre-processing programs

Modern network environment is sophisticated. Raw data that comes from such unstable surroundings may contain some unexpected contents. For example, data from website may somehow comprise HTML tags or some other strange symbols that do not belong to Unicode. Moreover, character is the most basic unit in Chinese and space is not treated as an element. Nevertheless, in most European languages, each word is split by one space and single character usually is meaningless. Such divergence leads to a word splitting problem. The pre-process program should solve all of them.

1.1.3 Achieving self-designed sentence alignment algorithm

Sentence alignment is the kernel for bilingual corpus construction. Aligned document cannot directly input into machine translation model. Only if each sentence of one document in language A has its corresponding one in language B, the corpus will be available. Hence, the sentence alignment algorithm should be able to precisely match two sentence from different languages that have exactly same meaning.

1.1.4 Developing evaluation metrics for both sentence alignment and translation

This task involves two kinds of evaluation metric, one for estimating the accuracy and performance of the sentence alignment algorithm, the other one for roughly assessing the translation quality of crawled corpus.

1.1.5 Constructing a cloud platform for manually modification

Corpus construction is not the only business for computer, but also human. Program can reduce the amount of labour and make the whole process efficient, but may not always generate correct answers. Thus, it is necessary to design a workplace for human. The platform is supposed to provide the function of corpus visualization, corpus information checking, corpus modification assistant, etc. It is also a platform that assembles every useful tool created from previous tasks.

1.2 Risk Assessment

Probability	High			Risk 1
	Medium		Risk 3	
	Low	Risk 2		
		Low	Medium	High
			Impact	

Figure 1: Probability impact matrix before proposed solution

Table 1: Table of prioritized risk

Priority	Risk Identifier and Description
1	Risk 1: When crawl data from the Internet, some websites prevent the crawler, so that it is hard to find enough corpora.
2	Risk 2: Unexpected cases occur in the cleaning process, which leads to the cleaning program crashed.
3	Risk 3: Unexpected problems occur on the local server.

Risk 1: When crawl data from the Internet, some websites prevent the crawler, so that it is hard to find enough corpus.

Description:

Crawling data from websites will burden their server. Therefore, many web servers will set up firewalls to prevent large stream from crawlers. As the result, the crawler program cannot gain much web information from them.

Solution:

Try to slow down the crawl rate and mask as a normal visitor. If still get prevented, find more website instead. Always keep substitute websites.

Risk 2: Unexpected cases occur in the cleaning process, which leads to the cleaning program crashed.

Description:

While programming the cleaning process, only cases that usually occur are under consideration. Thus, situations that have never been touched may exist. Those never-meet cases may cause something terrible on the whole program.

Solution:

Using try and exception in the proper place to prevent fatal incidents happening so that even though the unexpected case occurs, they will not ruin the entire program. However, this kind of problem cannot be totally avoided in this way, since it is unnecessary to consider all possible situations. What the cleaning program needs to do is only dealing with those common cases.

Risk 3: Unexpected problems occur on the local server.

Description:

The server running on the internet will suffer many kind of problems, such as vicious attack and high visiting rate. Hence, while building the server, precautions should be done to against most the future problem.

Solution:

Read more paper and books about server development. Try to prevent more problems before finishing setting up the server.

Probability

	Low	Medium	High
High			
Medium			Risk 1
Low	Risk 3	Risk 2	

Impact

Figure 2: Re-Assess impact/probability for each risk

1.3 Summary

In summary, the organization of this report includes three core topics: architecture design, algorithm design and platform design. The generalization for each chapter is listed below:

- Chapter 2 mainly illustrates the related subject and knowledge that the project involves. They can be roughly concluded in three levels: term level, algorithm level and research level. The term level primarily states basic terms which are involved in the project. The algorithm level explains each famous algorithm that has been used to solve the similar problem in detail. In the end, the research level introduces many related researches and familiar works which have been finished by others.
- Chapter 3 will describe the overall architecture of the corpus construction procedures. In this chapter, UML diagram is the commonly used manner. Besides, flowcharts and architecture diagrams are also imported to ensure the explanation is clear. There are several core components in this project, which are crawler server, data cleaning program, evaluation program and cloud platform.
- Chapter 4 shows the implementation detail with screen capture of codes. As this project involves in algorithm design, together with codes the design thinking will also be included in this chapter. In conclusion, there are three kinds of thinking, thinking of how to reduce codes and enhance the readability of the program, thinking of how to

enhance the performance of the program and thinking of how to fulfil the algorithm target.

- Chapter 5 presents the result and corresponding discussion into two varieties. The first one is concerning implementation outcome of web crawler, data cleaning and cloud platform. The second one analyses the algorithm result corresponding to sentence alignment algorithm and evaluation metrics.
- Chapter 6 concludes the whole project, meanwhile specifies the value of the corpus construction procedure and the future work.

2 Background and Related Work

What the project have done has the closest relationship with bilingual corpus construction and sentence alignment. Although these two tasks mainly related to Machine Translation [5], in general, there are two subjects that cannot be separated from it, Corpus linguistics [6] and Natural Language Processing [7].

This chapter divides all the related knowledge into three levels: term level, algorithm level and research level. The term level primarily states basic terms that are involved in the project. The algorithm level explains several famous algorithms that has been used to solve the similar problem in detail. The research level introduces many related researches and familiar works which have been finished by others.

2.1 Basic Terms

2.1.1 Corpus & Corpus Linguistics

To do NLP, there is an indispensable tool called corpus. The corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.[8] In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts (nowadays usually electronically stored and processed) [9]. Corpus linguistics studies linguistics based on Corpus.

Many technological developed country fund lots of money to build their own corpus such as British National Corpus. In China, Peking University Corpus, including modern Chinese, ancient Chinese and Chinese-English corpus, have been built since 1992. So far, it is one of the most influential Chinese corpora.

Ding Xinshan [2] stated from 1981 to 1991, there are 480 researches which base on corpus sponsored. With the enhancement of the fund, corpus building develops a lot. The second generation of the corpus has been established during 80s of last century [10]. Equipped with intelligence tools, they are able to process and edit corpus more efficiently than ever before. This project also aims at constructing the second generation of corpus.

2.1.2 Natural Language Processing

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages.

Thus, it is obvious that NLP processes the information which human generates other than devices. It is feasible for people to analyze such information, but as for computer it will be extremely difficult. However, only machine can deal with a large amount of text. Thus, the reason why NLP exists is to make text analyzable.

2.1.3 Corpus Construction

Europarl [11] is a multilingual corpus which collects up to 60 million words per language until 2011. The paper published in MT Summit 2005 has mentioned a corpus collection procedure that they used. In terms of parallel corpus construction, the very first step is to obtain raw data with the help of crawling technique. Then extract and map parallel chunks of texts in each language. This step is the so-called document alignment. After that, documents must be split into sentences (especially for European languages). The fourth step is to transform the data into proper form so that the machine translation (MT) system can directly use. Finally, map sentences in one language to the other language. This step is the so-called sentence alignment.

This project attempts to solve the foremost problem of bilingual corpus construction, especially for Chinese-Portuguese translation pair.

2.1.4 Statistical Machine Translation

Statistical Machine Translation (SMT) is a machine translation paradigm that depends on statistical models. The parameters that the statistical model used are generated from well-organized bilingual corpus. Statistical machine translation is related to other data-driven methods in machine translation, such as the earlier work on example-based machine translation. [12] This project obeys the data cleaning standard of Moses, a very famous SMT system. [13]

2.1.5 Sentence Alignment

Sentence alignment is a basic term in both bilingual corpus construction and machine translation. Gale and Church [14] defined that sentence alignment is a task to identify the correspondence between sentences in one language and sentences in the other language. There are two core methods that are used to align sentence, one is length-based and the other one is lexicon-based. Detailed knowledge will be mentioned below (2.2.2 & 2.3.2).

2.2 Basic Algorithms

2.2.1 IBM Model (1, 2)

IBM translation models form the basis for many SMT models used today. There are five IBM models. From the first one to the last, each model enhances the translation accuracy but exponentially augment the computational complexity. Besides, IBM models are the guidance of corpus collection. To know what kind of data is valuable for SMT, it is necessary to refer such fundamental knowledge. For the reason that IBM model 1 and 2 is the basis of five IBM model, here only briefly introduces them.

In the lecture of Micheal Collins [15], it is obviously declared that IBM models are an instance of noise-channel approach. As such they have two essential components, language model and translation model. Language model tells the possibility of the occurrence of a single language sentence. Assume the language is Chinese. Then the possibility will be $p(c)$, where $c = c_1, c_2, \dots . c_n$. Translation model stands for a conditional possibility. If translate Chinese into Portuguese, the symbol can be written as $p(pt|c)$. Sum up, given these two possibilities, the noise-channel approach is $pt_{best} = arg\ max_{c \in C} p(c) p(pt|c)$. The capital C in the formula represents the set (corpus) of all sentences in Chinese, while the lower case c is one of the sentences in Chinese. The formula can be explained in human word like that— how likely the Portuguese sentence pt is the translation of Chinese sentence c, under the prior of Chinese sentence c occurs.

IBM model 1 adds alignment to the original formula to reduce the difficulty of modelling. After converting, the conditional possibility becomes as follows:

$$p(pt_1 \dots pt_m, a_l \dots a_m | c_1 \dots c_l, m) \dots\dots\dots\dots\dots (1)$$

Where: m is the length of the Portuguese sentence, while l is the length of the Chinese sentence.

In this possibility, each pt stands for a word in the Portuguese sentence. For each Portuguese word, they correspond to zero or more Chinese word c_l at position a_l. However, the computational complexity has exponentially increased in this way.

$$p(pt_1 \dots pt_m | c_1 \dots c_l) = \sum_{a_1=0}^{l} \dots \sum_{a_m=0}^{l} p(pt_1 \dots pt_m, a_l \dots a_m | c_1 \dots c_l) \text{ .... (2)}$$

Thus, IBM model 2 is born.

IBM model 2 introduces a word alignment possibility $t(pt|c)$ where $pt \in PT$ and c $\in$ CN (PT represents all Portuguese words while CN represents all Chinese words) and possibility $q(j|i,l,m)$ where j is the word position in source language, i is the word position in target language, l is the length of source language and m is the length of target language.

The whole formula is transformed into this shape:

$$p(pt_1 \dots pt_m, a_l \dots a_m | c_1 \dots c_l, m) = \prod_{i=1}^{m} q(j|i,l,m) t(pt_i | c_{a_i}) \text{............. (3)}$$

With logarithmic formula, it can be converted into:

$$\prod_{i=1}^{m} q(j|i,l,m) t(pt_i | c_{a_i}) = \sum_{i=1}^{m} \log(q(j|i,l,m) t(pt_i | c_{a_i})) \text{............. (4)}$$

As a result, the calculation significantly decreased.

It is evident that sentence of both languages is the basic element of IBM model which is the same for SMT model. Besides, the training efficiency is directly related to the length of each sentence. These two conclusions are the guidance for the corpus collection and construction processes in the project.

2.2.2 Length-based Sentence Alignment Algorithm from Gale & Church

In 1991, Gale and Church [13] have already achieved a length-based sentence alignment algorithm. Up to now, many organizations still utilize their algorithm to align sentence in different situation, such as Europarl. The only sentence alignment algorithm that NLTK [16] provides is Gale and Church’s method. In their paper, they find sentence from one language has a very high possibility corresponding to sentence with similar length in the other language.

Gale and Church algorithm has two processes. The first step is paragraph alignment. Because paragraph length is highly correlated, paragraph should be directly one-by- one corresponded. However, this conclusion from Gale and Church may not always suitable in every situation. In fact, document aligned bilingual text in the modern sophisticated Internet may not always obey this rule. The paragraph amount of two corresponding documents may not be same.

After paragraph aligned, they use two parameters c and s^2 to measure the distance of sentence in two languages. c is the expected number of characters in l_2 (language 2) per character in l_1 (language 1), and s^2 is the variance of number of characters in l_2 per character in l_1. Both c and s^2 can be calculated from the documents. The standard normal distribution δ, whose mean is c and variance is s^2, equals to $(l_2 - l_1\mathrm{c})/\sqrt{l_1 s^2}$. The probability of δ can be calculated by the standard normal distribution table.

They then find out six matching conditions. They are one to zero, zero to one, one to one, one to two, two to one and two to two correlation. Based on these conditions, dynamic programming has been introduced to reduce the computation. The status of dynamic programming is $\mathrm{D(i, j)}$, where i is the sentence number from source language and j is the sentence number from target language. The status transform function is:

$$D(i,j) = \min \begin{cases} D(i,j-1) & + \quad d(0,t_j;0,0) \\ D(i-1,j) & + \quad d(s_i,0;0,0) \\ D(i-1,j-1) & + \quad d(s_i,t_j;0,0) \\ D(i-1,j-2) & + \quad d(s_i,t_j;0,t_{j-1}) \\ D(i-2,j-1) & + \quad d(s_i,t_j;s_{i-1},0) \\ D(i-2,j-2) & + \quad d(s_i,t_j;s_{i-1},t_{j-1}) \end{cases} \quad \ldots\ldots\ldots\ldots. (5)$$

Where d is the distance that has been worked out by c and s^2.

There are several special features of Gale and Church's algorithm that may perform different for other languages. The sentence length in their algorithm is the number of characters. However, the meaning of characters in European languages is extremely different from characters in Chinese and many Asian languages such as Japanese. Thus, if align English with Chinese, their algorithm cannot be directly used, otherwise the accuracy will be distinguished. Furthermore, documents must be split into sentence at the very first stage. Same as the bias in character, the concept of Chinese sentence is quite distinguished with European languages, which causes one Chinese sentence in most situation will correspond to more than one English sentence if split only by full stop. Thus, the match situation cannot be easily classified into six (see above). Because of these two factors, Gale and Church sentence alignment algorithm will not perform better on Asian language, especially Chinese.

2.2.3 Translation Evaluation Methods

Traditional machine translation evaluation fully relied on human experts consumes months to finish and involve human labour that cannot be reused. Human evaluations of machine translation (MT) weigh many aspects of translation, including adequacy, fidelity, and fluency of the translation. [28]

- Bilingual Evaluation Understudy (BLEU) Score [17]
  Researchers in IBM have released a semi-automatic translation evaluation metric, called BLEU score. Together with machine translation result, it needs the human translation to be the reference. Therefore, it is a semi-automatic evaluation method. N-gram precision and brevity penalty are two parts of BLEU. Precision itself is a commonly used score for many statistical models and machine learning models. In terms of machine translation evaluation, precision reflects the translation sufficiency only. It is a fraction of the number of word of candidate sentence that occurs in the reference over the total number of words in the candidate sentence. Besides sufficiency, fluency is also necessary. Thus, n-gram is added into the precision so that not only count each single word, but also two, three and more adjacent words. The final precision score is the geometric mean of n-gram precision:

$$p_n = \frac{\sum_{w \in candidates}^{C} \sum_{n-gram}^{N} Count_{clip}(n-gram)}{\sum_{w \in candidates}^{C} \sum_{n-gram}^{N} Count(n-gram)} \ldots\ldots\ldots\ldots\ldots (6)$$

  After geometric average n-gram precision becomes:

$$\sum_{n=1}^{N} W_n \log p_n \text{…………. (7)}$$

Where W_n is always $\frac{1}{n-gram}$ for each gram.

Nevertheless, only n-gram precision is not enough. It cannot detect the situation if the number of word in candidates is far from enough. Sentence brevity penalty is the punishment score which will be bigger is the length of the candidates is less than the reference.

$$\mathrm{BP} = \begin{Bmatrix} 1 & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \le r \end{Bmatrix} \text{…………. (8)}$$

In which c is the length of candidate sentence, while r is the length of reference sentences.

Thus, BLEU score is written as:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} W_n \log p_n\right) \text{ …………. (9)}$$

The range of the score is between 0 and 1, in which 1 means all the word in n-gram evaluation scale that the candidate sentence comprises appear in reference sentences. Usually the reference sentences are at least two varieties of translation.

In conclusion, BLEU score regards the reference sentences as the evaluation benchmark. However, even human now cannot say, given one sentence, there is only one best or accurate translation. Thus, the authority of reference is the basic factor of judging whether the translation is good or bad. Besides, the semi-automatic evaluation metric still demands lots of human resources, which will be a fatal obstacle for many evaluation cases, such as estimate the translation quality of the bilingual text in webpages.

- mWER[18]

  Nießen S proposes an evaluation method based on Levenshtein distance [19]. Similar to BLEU score, mWER also needs reference sentences as the benchmark, but it will consider the Levenshtein distance as the score value. Levenshtein distance, also called edit distance, is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into another. It can be easily calculated with the help of dynamic programming. The status transform function is:

$$\mathrm{d(i,j)} = \begin{Bmatrix} \max(i,j) \\ min \begin{Bmatrix} d(i-1,j)+1 \\ d(i,j-1)+1 \\ d(i-1,j-1)+1 \end{Bmatrix} \end{Bmatrix} \text{…………. (10)}$$

Where $d(i-1,j)$ stands for insertion, $d(i,j-1)$ stands for deletion and $d(i-1,j-1)$ stands for substitution.

Using this distance measure algorithm, it is easy to find how big the difference is between two sentences. Therefore, this algorithm can be utilized on sentence alignment evaluation.

2.3 Related Researches

2.3.1 Bilingual Web Crawler

There are three major structures of parallel documents in the World Wide Web, parent page structure, sibling page structure, and monolingual sub-tree structure [20]. Parent page structure means there is a parent page which links to different language versions of the website. Sibling page structure refers to the case in which the page in one language contains a link directly to the translated pages in another language. Monolingual structure represents the web structure that totally separate different languages such like two distinguished websites. In terms of the first two structures, it is relatively easy to directly crawl data that have already been document aligned. Although sentence alignment is the target form, it is acceptable to crawl the document-aligned data at first.

Ma and Liberman [25] have developed a Bilingual Internet Text Search system, especially for the third kind of web structure to determine whether two documents are translated to each other or not. They use tokenization to calculate the similarity of two documents. If the score is higher than a pre-defined threshold, then those two documents will be treated as a pair.

In terms of speed, Europarl declares what they have done using web crawler. They spent several days to obtain 80,000 files each language, but compared with pure artificial translation it is satisfactory. As raw data obtainment does not mean everything accomplished, there are many procedures that needed to be done manually later. Thus, speed of crawler is not the central obstacle compared with other challenges when constructing corpus.

2.3.2 Various Solutions on Sentence Alignment

Apart from the respective contribution from Gale and Church, there are many remarkable researches on this topic.

Brown, Lai and Mercer [21]'s length-based sentence alignment algorithm is a little bit different from Gale and Church. They use word length to describe the sentence length. However, the result is not satisfactory. Besides, Gale and Church also attempt the same method. They concluded that characters are more than words therefore there is less uncertainty.

Lexicon-based sentence alignment algorithm is also welcomed. Warwick and Russel [22] use a bilingual dictionary to select word pairs in sentences from a parallel corpus and then align the sentence based on such bilingual dictionary. Moreover, Simard et al. [23] finds a type of lexicon which contains more information, called cognate. However, cognate only exists in languages with close history. Thus, his discovery is not popular used.

Sun et al. [24] combine the length-based approach with lexicon check. They design a score which is used to judge if the length-based alignment is correct. If the score is higher than a defined threshold, the alignment is correct. Then eliminate the correct part and do the length-based alignment. These procedures will be circulated until every sentence is removed.

Yang and Li [26] have successfully used sentence alignment technique on alignment of English and Chinese title. However, title alignment is exactly another story.

Li Yiping et al. incorporate Gale and Church's length-based sentence alignment algorithm and lexicon cues on Chinese-Portuguese sentence alignment. They have successfully obtained nearly one hundred correct results on their test corpus. Whereas, they only give some examples of the lexicon cues, and the cue only serves for Chinese-Portuguese pair. Thus, the expandability is not good.

Wang Fei [27] also combines length-based and lexicon-based methods together. He effectively uses lexicon cue to find anchor point. Anchor point is a gap that can shrink the pending content into smaller one. This process is same as doing further alignment before sentence alignment and after paragraph alignment. Although, his result is better than some other sentence alignment algorithms, there is a fatal problem of his method. If the anchor point is wrongly set, the further result under this anchor point will be entirely affected. Thus, the robustness is not good.

In conclusion, researches that only use length-based approach have worse accuracy but run faster than those algorithms with combination approach. Some of methods are not appropriate for Chinese sentence, while others only work on Chinese corpus. There is one important issue that nearly no research tends to discuss. The Chinese sentence may be very long in a long document, especially in a literature article. In prediction, a long sentence will match more than one, even two, sentence of other languages. In terms of extreme situations, one paragraph is possible to be one long sentence in Chinese. Therefore, if the algorithm only treats sentence as the smallest unit, there must be lots of one-to-many correspondences. In consequence, the alignment will be meaningless. To solve such special problem, it is pivotal to use other smaller units instead. Therefore, the algorithm in this project utilizes comma as a better split symbol.

3 Methodology

To complete this project, there are five components, which are corpus collection, data cleaning, algorithm design, evaluation metric design and cloud platform design, to be done. Respectively, five sections are listed below. Here is the overall structure of the project:

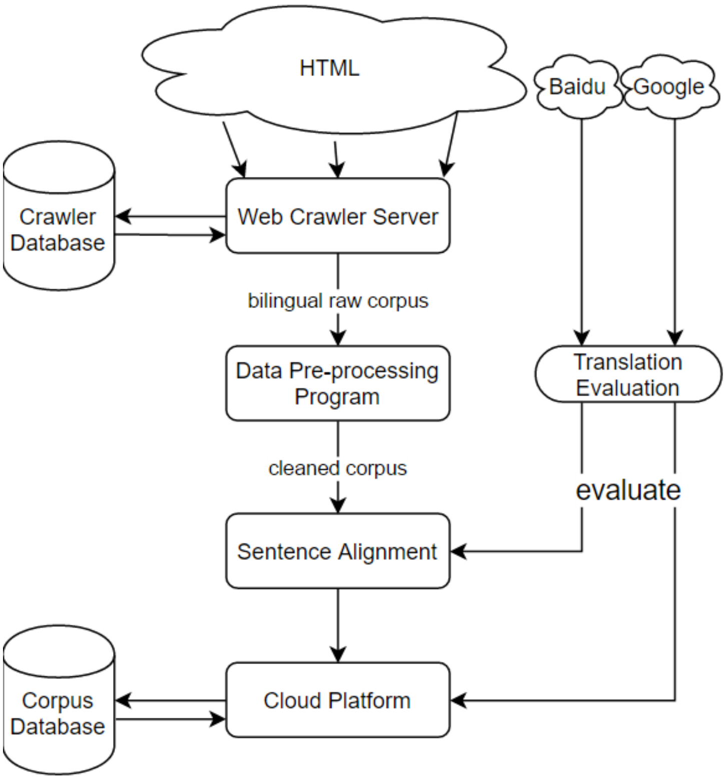


Figure 3: Overall project structure

- In the upper most section, it will describe the mechanism of bilingual web crawler and three kinds of web structure that the crawler overcomes. Besides, the crawler server architecture will be shown and explained.
- After obtaining the output from web crawler, there will be a python program which is dedicated to cleaning the data. Data cleaning contains some steps, thus, the organization of each step and their functions will also be illustrated.
- The third component is the sentence alignment algorithm. There are three procedures in the algorithm, which are paragraph alignment, long sentence alignment and short sentence splitting. In this section, it will explain the reason why each step is required and how they work.

- The fourth part comprises the translation evaluation metric. It aims at roughly estimating the translation quality in a full-automatic way. This section will indicate the meaning of each evaluation.
- Last but not least, the fifth section is a cloud web design. Together with the design ideas, it will reveal and explain two central diagrams that the development needs, ER-diagram and sequence diagram.

3.1 Bilingual Corpus Crawler

The bilingual crawling skill of three kinds of web structure will be shown respectively. Besides, all of them are assembled on one crawler server. The architecture of the crawler server will also be shown.

3.1.1 Crawler of Website in Macao

Macao government website has abundant Chinese-Portuguese bilingual contents. They are the best targets to collect bilingual corpora. Their web structure (This is called sibling web structure) is the most convenient class to crawl. The structure diagram likes:

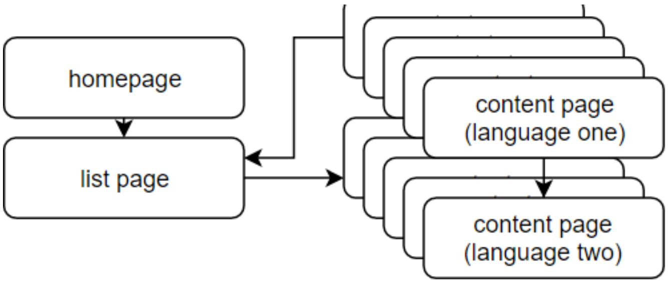


Figure 4: Sibling bilingual web structure

To some extent, homepage for mass of website is only a portal, which is useless to crawl. It is central to find the list page hidden in the homepage. List page usually directly link to the interface to fetch data from the website database. Thus, in general, it is treated as a content container. For example, in www.io.gov.mo , the hyperlink of list page is on the left-hand side of the homepage. Depends on the website design, list page usually may not be single. Multi-list page leads to two circumstances as follows:

- Different list page uses unequal HTML design.
- Each list page has exactly same HTML structure.

If the first one happens, it is inevitable to treat the list page as another website. Therefore, another script is demanded. For the second situation, only thing need to be done is adding one level on the crawler hierarchy, which means regarding the homepage as the first list page and regarding the original list page as content page like the following structure:

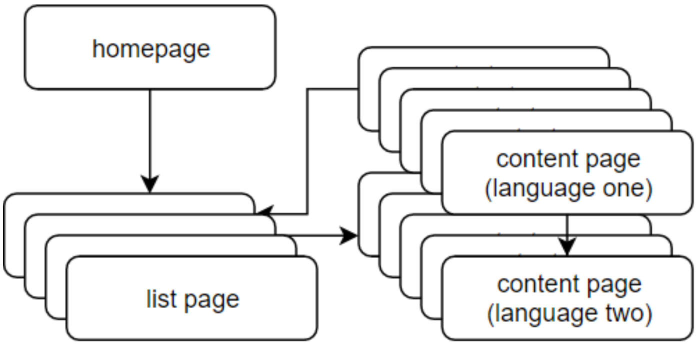


Figure 5: Multi-level sibling structure

Content pages are the target page that the crawler needs to find. In most situations, for the sibling web structure, most of the content pages have indirect link on the list page. Because all of the content links cannot emerge on the list page at the same time, only one category of content will be shown when the users first enter the list page. As a result, drop down lists become the common tool to selectively display the content.

Each content page, in bilingual sibling structure, must contains one link that straight points to the other language. The most convenient feature is that those contents should have been document aligned already.

3.1.2 Crawler of TED

TED is a typical representative of parent bilingual web structure. There is a pivot page that directly links multi-language TED speeches. The structure diagram likes below:

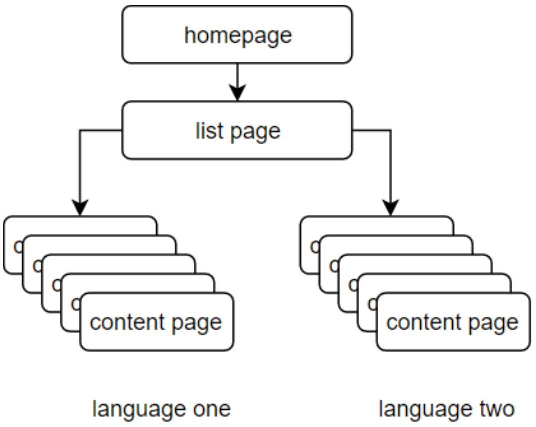


Figure 6: Parent bilingual web structure

The character of this web structure is that there are no links which give the portal for content pages to directly jump to different language version. In order to crawl this kind of website, it is possible to record the page with a unique variable. However, the project uses another method that entirely crawl all valuable information of TED. The crawling flow chart is as follows:

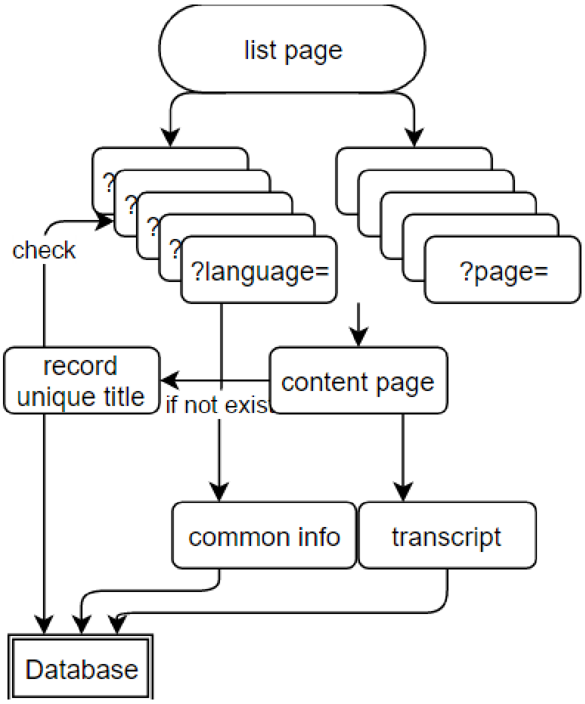


Figure 7: TED exhaustive crawling structure

By analysing the URL of TED list pages, it is obvious that each page contains the two key queries, “language” and “page”. “language” indicates the language that the page lists, while “page” presents the current page number. Using these two queries, JavaScript operation such as selecting the drop-down list items or clicking buttons can be escaped. In conclusion, the first level focus on finding required content page. After obtaining the link of content page, three significant items can be easily discovered, which are unique title name, common information and transcripts.

- Unique title name is an English string that combines both author name and speech title. It is used to identify which speeches have been crawled.
- Common information, such as topics, rates and comments, is same for each speech with identical title name, even though the language of speech is different. Therefore, it should only be crawled once.
- Transcript is the subtitle of TED speech which has multi-language versions. Hence, in terms of speech with same title name, it still needs to be crawled many times until all language versions have been retrieved.

Based on the above crawling structure, the ER diagram is as follows:

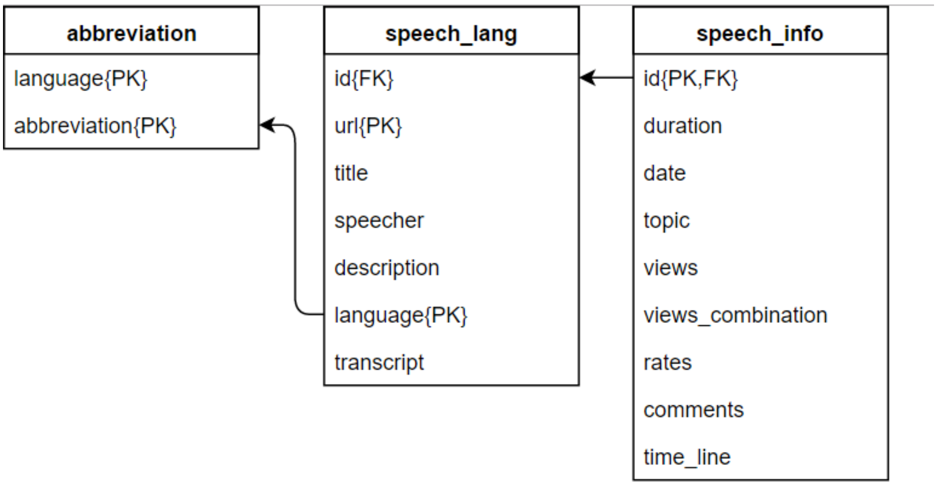


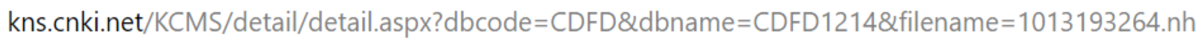
Figure 8: ER diagram of TED data

The corresponding data structure is more complicated than Macao websites. There are two tables. The “speech_info” table stores the unique information of each speech. The “speech_lang” stores the language information that is distinguished depending on different languages.

3.1.3 Crawler of CNKI

To make the corpus quality as high as possible, CHKI (www.cnki.net) is the best choose. Nearly every paper which can be found in this website is highly qualified. However, the high security level prevents crawlers from obtaining data casually. For example, the website will suspect and block the crawler, if it crawls the data quite faster than human access. Moreover, to protect the copyright, the website only public the abstract of paper and block the other contents of each paper, which means, only abstract can be crawled. Good news is that not only the abstract that can be used, the published date, author, download frequency, reference frequency, type of each paper can also be exploited. Although in this project they are useless, these worthy data must be helpful in data mining in other fields. Thus, those data are also considered as the target.

As CNKI has nearly invulnerable crawler prevention mechanism, URL is the only weakness that can be invaded. For instance, see the URL below:



kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CDFD&dbname=CDFD1214&filename=1013193264.nh

The parameter "filename" has a long integer as the value. It is possible to directly use such kind of URL to enter the content page. Whereas, there may be some numbers corresponds to invalid addresses. In fact, only a few of them are valid. Besides, it is not easy to find the rule behind the number only by observation. One possible idea is to exhaustively crawl every URL. In this way, by roughly estimating, three ten thousand points need to be extracted from ten million nodes. To shrink the search scale, this project basically performs statistical analysis on the URL to find relation among those valid numbers and invalid numbers.

Here comes the statistical analysis and the program architecture:

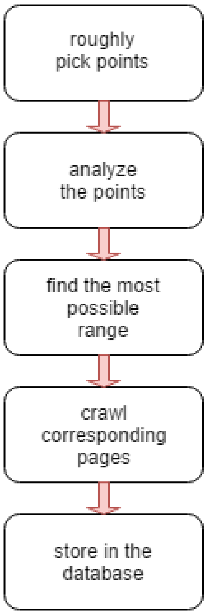


Figure 9: CNKI URL analysis processes

The first step is to randomly pick several points from the ten million points to roughly find the distribution of the node. In the second step, all the result comes from the first step will be put into MATLAB for data analysis. After finding the most possible range, it is time to start the real crawler. In the end, preserve all the crawled data into database.

3.1.4 Crawler Server Architecture

Crawler should automatically crawl each website, otherwise manually starting those crawler will be still troublesome. There are six website mentioned above. Hence, at least six web crawler scripts are needed to run on the server. According to the web structure, those six scripts can be divided into 3 varieties, parent structure, sibling structure and CNKI. A schedule is designed to automatically run each crawler on required time. Besides, in terms of recrawling the failure URL, a stack is set for collecting the failure ones. Furthermore, each crawler executes simultaneous so that they will not influence each other.

This is the crawler server architecure diagram:

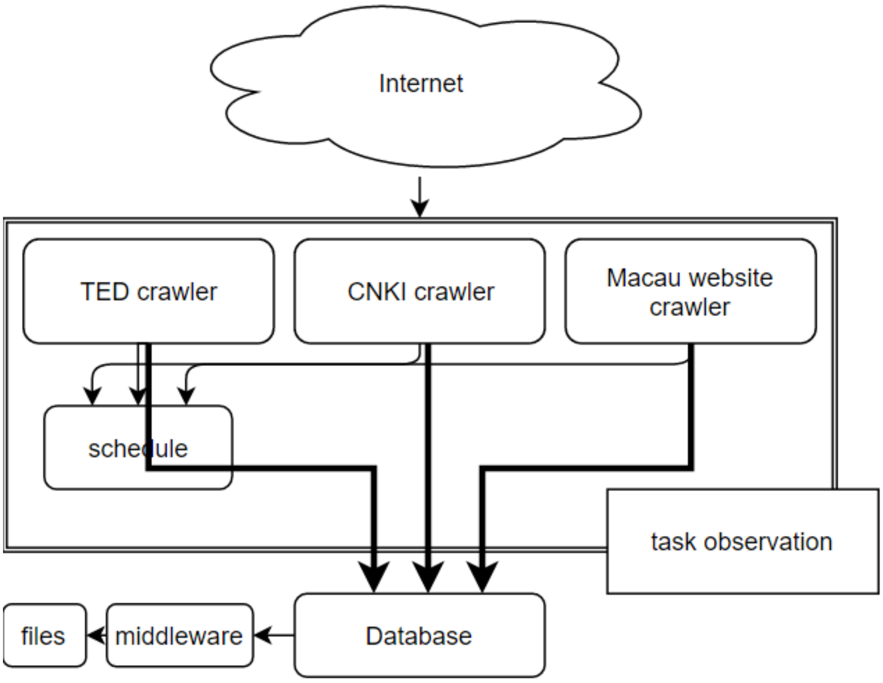


Figure 10: Crawler server architecture

There are five components in the server and each of them runs individually. The upper three are the crawler programs. They not only have to crawl required information from website, but also submit the task process to the schedule program. The schedule program needs to fulfil two objectives. One is to judge whether it is the time to restart the crawler or not. The other one is to record failure tasks, and rerun them after the accomplishment of all active tasks. If every step is successfully executing, results will directly be preserved into Databases. Due to the function of schedule module, there is a task observation interface which helps user monitor the progress. Moreover, if files output is demanded, the middleware is ready to select the proper data and convert them to text format or any other formats.

3.2 Data Pre-processing Program

Data pre-processing which is another key step in corpus construction mainly focuses on filtering useless contents and tokenization, which will be explained in this subsection in sequence.

3.2.1 Tokenization and Sentence Splitting

Word splitting for European language, such as English, is called tokenization. Tokenization does not exist in Chinese and other Asian languages which do not use space as division.

The period in most European language usually acts as the end of a sentence. Nevertheless, it also has other functions. For instance, period is the symbol that indicates abbreviation. The period can also be the break of URI (web address). Thus, the tokenization is supposed to detect different usage of periods.

Furthermore, same as period, comma and space in some situation have ambiguity as well. For example, the number 6000 can be written as 6,000. In this form, comma is only a mark that specifies a thousand.

Here is a diagram that shows each function in the tokenization and their relationships:

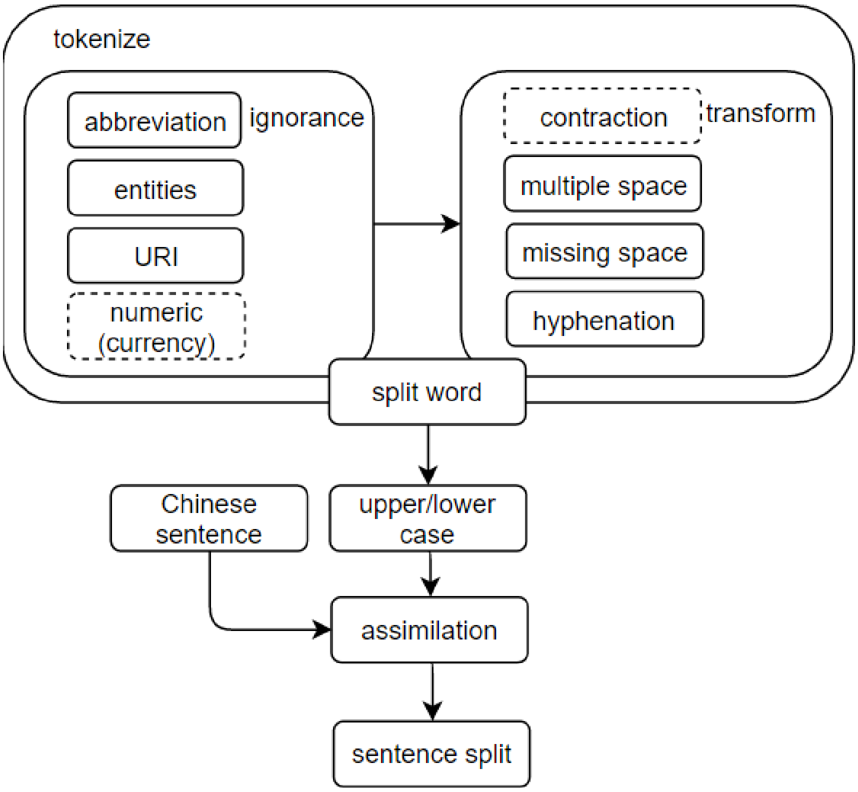


Figure 11: Tokenization

Firstly, it is essential to list out the items that need to be ignored. Abbreviations like Feb., Entities like “&eacute;”, URIs like “www.baidu.com”, numeric units like “10kg” are required to be accurately detected. There is one exception. In Penn Treebank [28], units except “%” are not allowed to be split. Thus, it is not supposed to split “10kg” into “10 kg” but should split “10%” into “10 %”. The ignorance is the case that should not be separated. The next step is to transform some item combinations into the required ones. Multi-space likes “I am” should be transformed into “I am”. Missing space likes “right,let’s go.” should be converted to “right, let’s go.” Hyphenation likes “dic- \n tionary” should be changed to “dictionary \n” (“\n” stands for line feed). Similarly, contraction in the Penn Treebank is not transformed.

After those procedures, the ambiguity of word will be eliminated. The process which is located below the diagram is called standardization. They unite the case form and the punctuation form. For instance, the Chinese punctuation “。” should be standardized into the European language format “.”

When do sentence splitting, it is not rigorous that only split sentence by full stop, exclamation mark and question mark. Besides them, marks which have open and close balance, such as quotation mark and brackets, should also be taken into consideration.

3.2.2 Data Filtering

HTML tags and duplicated sentences are the objects needs to be filtered. Using regular expression, HTML tags can be easily detected and removed. However, detecting duplicated sentence is relatively a hard work. Firstly, sentence should be split before deletion (previous step). Then hash them into a dictionary (Python). It is unnecessary to hash the whole sentence, instead only hashing some of the front word in the sentence is also feasible. This program is designed to calculate the average sentence amount in the document. Half of the average is treated as the hash amount. Nevertheless, duplicated sentences will be eliminated only in document level rather than corpus level.

3.3 Sentence Alignment Algorithm

The sentence alignment algorithm bases on the length of the sentence to intelligently find corresponding sentence pairs. The difficulty of length-based sentence alignment algorithm comes from the different sentence structure of two languages. For example, comma is not universally used in English-like language, but comma appears in nearly every Chinese sentence. The sentence structure problem prevents all the static align rules such as divide sentence roughly by punctuations. Whereas, only using dynamic aligning rule is still far from enough. There are three procedures in the algorithm, which are paragraph alignment, full-sentence alignment and sub-sentence alignment. Here is a flow chart shown the procedures in detail:

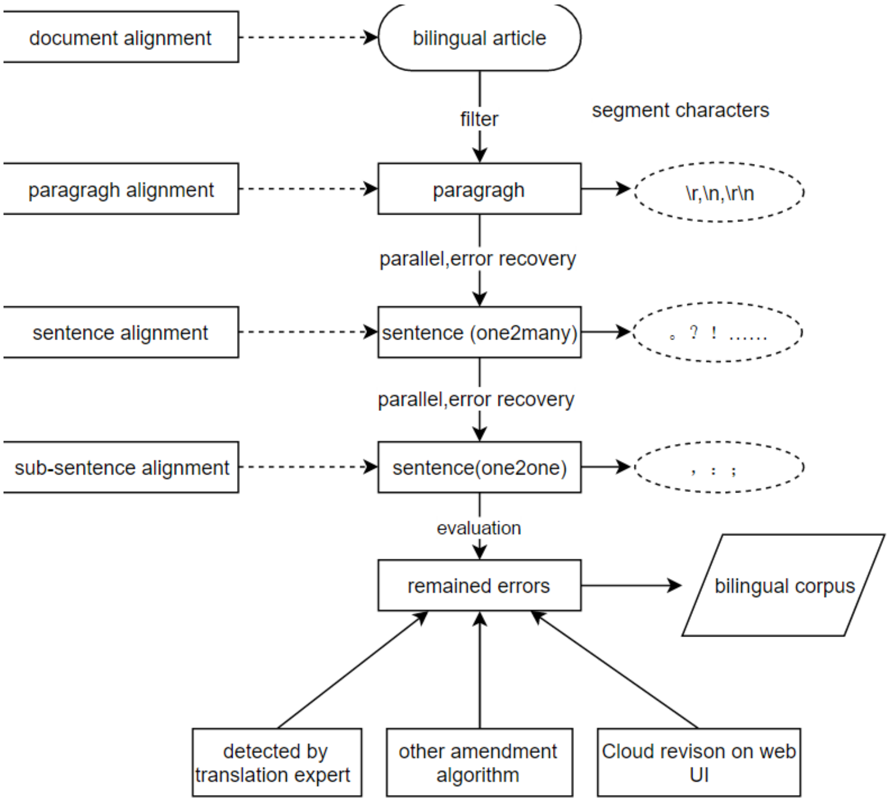


Figure 12: the flowchart of Chinese sentence alignment algorithm

The basic idea of the flowchart is smashing the whole parallel articles into smaller parallel units. The parallel units can be paragraph, full-sentence and sub-sentence. The algorithm performs iteration on each unit until every unit is successfully aligned. The input text must have already been document aligned. There is only one parameter called length rate, which means the ratio of the length of source sentence over the length of target sentence length:

$$\text{Length Rate} = \frac{source\ sentence\ length}{target\ sentence\ length} \ldots\ldots\ldots\ldots\ldots (11)$$

The calculation of length rate will be frequently used.

$$\mathrm{R} = \frac{\sum Cn{-}array}{\sum Pt{-}array} \ldots\ldots\ldots\ldots\ldots (12)$$

Assuming Chinese and Portuguese as the example, the capital R is generated from the sum of Chinese sentence length over the sum of Portuguese sentence length. Besides, the R will be the baseline in further steps.

The overall alignment matrix is as below:

$$Alignment = \begin{bmatrix} A_{s1} & A_{t1} & R_1 \\ A_{s2} & A_{t2} & R_2 \\ \vdots & \vdots & \vdots \\ A_{si} & A_{ti} & R_i \end{bmatrix} \ldots\ldots\ldots\ldots\ldots (13)$$

Where A_{si} stands for the amount of sentence of source language in the i-th alignment. Similarly, A_{ti} stands for the amount of sentence of source language in the i-th alignment. Correspondingly, R is the sentence length ratio of source divided by the sentence length ratio of target.

The alignment matrix will be initiated from one-to-one alignment for each alignment. However, in most situations, the number of target sentence is not the same as the number of source sentence. Thus, over-flowing sentence will be listed as many-to-zero. The algorithm will adjust it in later steps. After initiation, the relative ratio can also be calculated out. The largest ratio will be tinkered first.

For each adjustment, there are totally four possible cases that may improve the ratio values.

$$\text{cases} = \begin{Bmatrix} A_{bigger,i} - 1, A_{bigger,i-1} + 1 \\ A_{bigger,i} - 1, A_{bigger,i+1} + 1 \\ A_{lower,i} + 1, A_{lower,i-1} - 1 \\ A_{lower,i} + 1, A_{lower,i+1} - 1 \end{Bmatrix} \ldots\ldots\ldots\ldots\ldots (14)$$

Pick the case that makes the average of the influenced R_i close to R. In real cases, it is unnecessary to trace back, so at most 3 cases for each movement. When judging the next movement, pick the movement that causes a better R_i. For example, assume the previous R_2 equals to 0.5. After the movement A, the current R_2 becomes 0.3. After the movement B, the current R_2 becomes -0.9. Compare the absolute value of each movement. It is clear that 0.3 is smaller than 0.9. Thus, choose the movement A first. If no movement can make the R_i be better, the search will be completed. Then search the second-best choice.

In the end, pick the alignment matrix whose sum of absolute value of R_i is the lowest. The formula is shown below:

$$Alignment_{best} = \text{argmin} \sum_i^N (|R_i - R|) \ldots\ldots\ldots\ldots. (15)$$

Where N is the total amount of the alignments and i is the number of each alignment.

The basic idea of this algorithm is to adjust the length rate for each sentence pair with different sentence combination. Thus, it will perform well on those translations whose length is closed related. It is indispensable that special cases may occur even though most translation pairs obey the rule. Meanwhile, the special case may make the algorithm confused and made wrong decisions.

However, even the algorithm is powerful, the quality of the ultimate result still highly depends on the quality of the origin data. The effectiveness analysis of the algorithm is given in Chapter 5 (5.1.3.3).

3.4 Evaluation Metrics

This part will introduce the algorithm in three sections: (1) Overall description, (2) Feasibility analysis and (3) Effectiveness analysis

3.4.1 Algorithm Description

BLEU is a semi-automatic machine translation evaluation metric which requires human translation as the reference. This project substitutes Google and Baidu translation for human translation so that the entire process of the translation evaluation becomes full-automatic. However, the drawback is obvious. Using another machine translation results as the reference will surely limit the accuracy of this evaluation, as it is difficult to ensure that such reference is authoritative.

In order to enhance the accuracy, this project uses a two-way evaluation method, direct reference and indirect reference.

- For each sentence pair, direct reference treats Chinese to Portuguese translation results as the reference, while the original Portuguese sentence is the hypothesis.

- Indirect reference uses English as a middle language for the reason that language to English machine translation technic is more mature than translations between other languages. As a result, Chinese will be translated into English first. Then the English output will be translated into Portuguese in the end.

After obtaining and organizing the references, BLEU score will be used to give an appreciable number which is between 0 and 1. The score will be higher, if the translation sentence is closer to the reference sentence. Given this two-way evaluation method, the ultimate results pick the bigger one. Thus, the formula can be written as:

$$BLEU_i = \max_i(BLEU_i(Direct), BLEU_i(Indirect)) \ldots\ldots\ldots\ldots\ldots (16)$$

$BLEU(Direct)$ and $BLEU(Indirect)$ correspond to those two methods mentioned above. The translation which BLEU score is less than a reasonable threshold will be marked as a controversial one.

Furthermore, not only for each sentence pair, each article should also have an overall evaluation score to indicate whether two documents are aligned correctly. The judgement method in this project is to calculate the sum of the approved pairs (BLEU score over 0.3). If the sum is greater than 90% of total sentence amount, then the document alignment is treated healthy.

3.4.2 Feasibility analysis

There are two vital obstacles in front of achieving the goal, how to get the translation reference from Google and Baidu and how to calculate their BLEU score.

- Feasibility on obtaining translation results:

Fully utilizing the Google and Baidu API is a straightforward method. Whereas, there are some restrictions when directly use them. Price is the first biggest problem. Baidu translation API requires 49 RMB per million characters [30], while Google translation API requires 20 USD per million characters [31]. However, the price limitation is difficult to avoid or break through only if using plenty of available accounts.

Therefore, this project rejects the usage of translation API of both Baidu and Google, instead it uses crawling technology to gain the translation results directly from the webpage. In this way, there is no limitation on getting the result. Nevertheless, more time-consuming will occur due to parsing the HTML elements from the response.

- Feasibility on calculating BLEU score:

Compared with rewrite a BLEU algorithm, a Python natural language processing library called NLTK provides an encapsulation of BLEU which is more functional and perfect. Thus, there is no need to re-implement such algorithm.

3.4.3 Effectiveness analysis

BLEU is a score with the range of 0 to 1. The inner feature is that if hypothetical sentence (sentence to be measured) is more similar to the reference sentences (sentence get from Google and Baidu), the score will be closer to 1. Similarly, despite of brevity penalty, 0.5 roughly means half of the elements, including units from 1-gram to 4-gram, occur in the reference sentences. Hence, if the reference sentences are the perfect translated sentences, the similarity of reference sentence and hypothetical sentence will directly mean that the translation result that is to be measured is high quality.

In terms of the number of reference, more reference sentences will also increase the score due to the appearance of the word that in the hypothesis sentence will be enlarged. Therefore, in some extent, more reasonable references make the evaluation result more robust.

With regard to two-way reference, it is difficult to judge whether direct reference or indirect reference can lead to better BLEU score. One thing that can be confirmed is that if either of them gets a better score, the hypothetical sentence will be potential to be better. Otherwise, two irrelative sentences will not get evident higher score only by changing the way of evaluation (direct and indirect). Therefore, the maximum of direct and indirect evaluation score is the better choice.

3.5 Cloud Platform

This section will mainly introduce the design ideas of this cloud platform. The basic target of the cloud platform is to build a friendly environment for human translators so that their work of constructing bilingual corpus will be more efficient. Moreover, this platform combines many corpus construction technologies, including data cleaning, sentence alignment, translation evaluation and so on. Some of them, such as data collection and data cleaning, will be executed in the back end. On the contrary, sentence alignment and translation evaluation will involve in both front end and back end. Detailed design ideas will be described with the help of both sequence diagram and ER diagram.

3.5.1 Sequence diagram

This sequence diagram illustrates the interaction among students (normal translators), teachers (expert translators) and the cloud system.

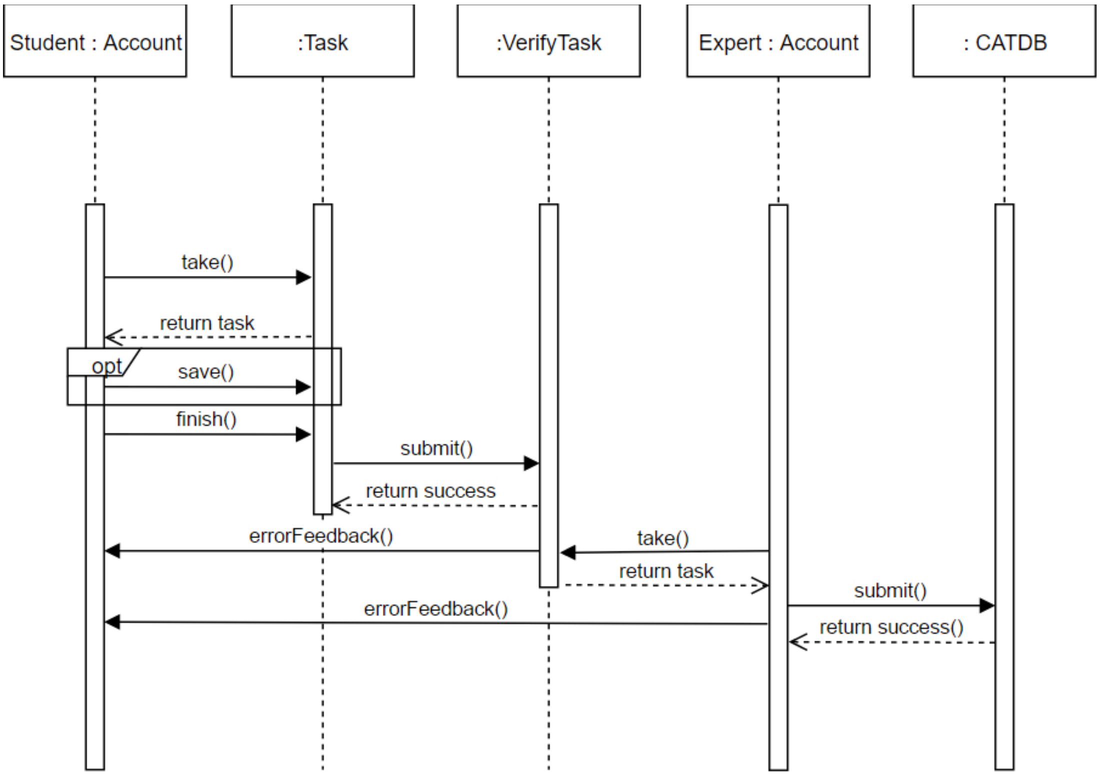


Figure 13: Sequence diagram of cloud platform

Students play the central role of the entire platform. They will do most of the initial modification of the corpus. Students can get a task from the system. Each modification is provided to students in the form of the task. After obtaining the task, the system will furnish a work platform for the task receiver. This work platform enables receiver not only to modify the sentence alignment mistakes and the translation mistakes, but also highlight the error reason so that researchers can reuse the errors to trace the algorithm problems. If the student cannot finish the task during this time, he or she can save the work or mark the sentence that is omitted at that time. The tasks, no matter accomplishing or not, will be store in the database waiting for verification or continuing. The task whose percentage is 100% will be automatically submitted to "VerifyTask" table. Experts are supposed to quickly skim the result deriving from students and amend their controversial answers. Experts must be reliable and accountable, because there is no secondary verification after expert confirmation. Experts can also feedback the error and remark the job of students. After experts' verification, the document will be written into the corpus. Administrator can view and modify everything that is in the corpus.

3.5.2 ER diagram

This ER diagram shows the entity and relationship of the database. As most of the design detail has been introduced in the previous section, this part only states something special.

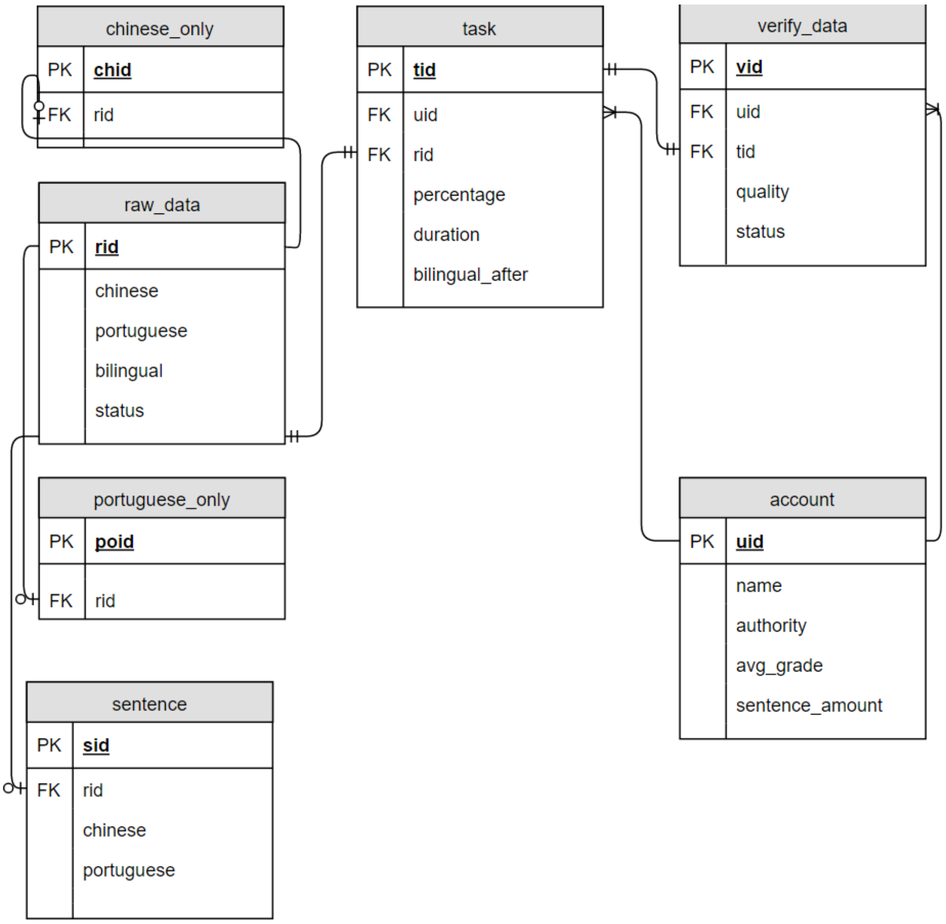


Figure 14: ER diagram of cloud platform

The central corpus comes from the table called "raw_data". It is also the destination of the data collection process. Chinese and Portuguese texts in each record are document aligned.

Each record has an attribute called status, which records whether this document is passed or not. If the document has not been processed, the status will be marked as “pending”. If the document has been processed and finished, the status will be marked as “success”. If the document has not correct aligned, the status will be marked as “cancelled”. “Chinese_only” and “Portuguese_only” are two tables used for indexing the document that only consists of single language. In “task” table, the “bilingual_after” is a json-type attribute with complex information inside. The json structure will be introduced in chapter 4.3 in detail.

Accounts comprise three authorities, which are the student, the teacher and the administrator. The avg_grade is one of the Key Competency Index (KCI), calculated by a simple formula:

$$Avg_{grade} = \frac{1}{N}\sum_{i}^{N}\frac{Amount_i - \alpha Error_i}{Day_i} \text{.............} \ (17)$$

Where N is the total number of tasks that the student completed, i is the task from 1 to N. $Amount_i$ is the total amount of sentence inside task i. $Error_i$ is the total amount of error that experts found out inside the task. Day_i is the duration from the beginning of the task to the accomplishment date. α is a penalty coefficient that indicates to what extent to treat the error. In this project, α is 2.

The “avg_grade” implies how many sentences the student can finish minus errors they made per day in average.

The “sentence_amount” is the other KCI that indicates how productive the student is.

With the help of KCI, it is easy to take care of those who are lazy or negligent and praise those who are efficient and aggressive.

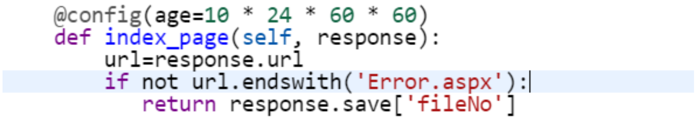
4 Implementation

The implementation includes three central sections, crawler, processing scripts and web implementation. Detailed explanation of programs together with the code capture will be introduced inside each section. Most statement and codes will be put in section 4.2, as this section is the kernel part of the whole system. Besides, there are two program language tables to illustrate the detailed languages that are used to achieve targets.

4.1 Crawler Implementation

Crawler implementation based on a Python crawler framework called pyspider. Combining with Phantomjs, the framework is functional and powerful. It uses multithread technique to accelerate the crawling speed. Meanwhile, with the help of Phantomjs, JQuery codes can be directly run together with Python grammar. Thus, avoiding the use of regular expression, it highly increases the readability of the program codes. Moreover, pyspider is a crawler server. It uses schedules to crawl each target website in required time intervals. More than one crawler can simultaneously run on it.

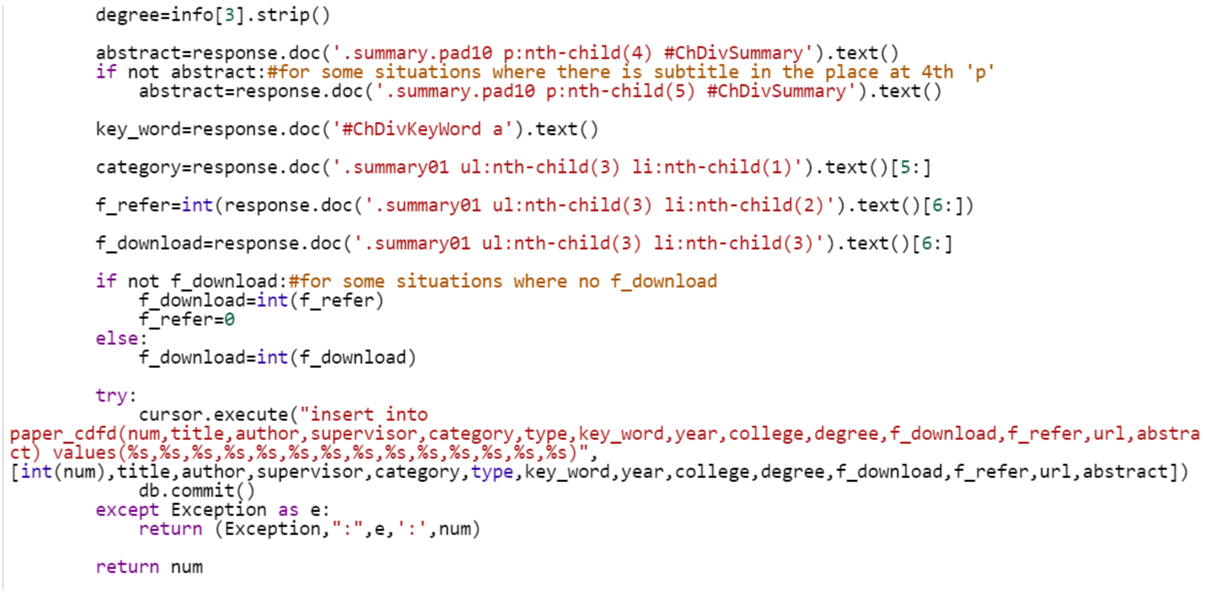
The following code captures show the distinctive part of the crawler program.



```
@config(age=10 * 24 * 60 * 60)
def index_page(self, response):
    url=response.url
    if not url.endswith('Error.aspx'):
       return response.save['fileNo']
```

Figure 15: Partial codes for CNKI URL detector

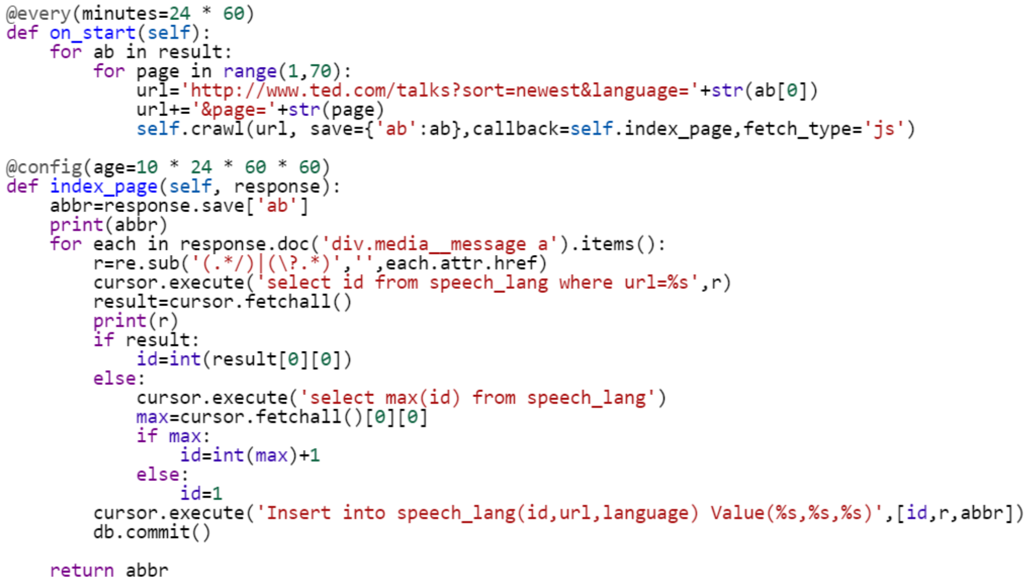
While crawling CNKI, URL should be firstly detected in order not to exhaustively search every possible node. This code will record those pages which contains the target contents. The search logic only seeks zero point one percentage of the whole nodes. After analysing the result, it is possible to obtain a range of valid nodes.



```
        degree=info[3].strip()

        abstract=response.doc('.summary.pad10 p:nth-child(4) #ChDivSummary').text()
        if not abstract:#for some situations where there is subtitle in the place at 4th 'p'
            abstract=response.doc('.summary.pad10 p:nth-child(5) #ChDivSummary').text()

        key_word=response.doc('#ChDivKeyWord a').text()

        category=response.doc('.summary01 ul:nth-child(3) li:nth-child(1)').text()[5:]

        f_refer=int(response.doc('.summary01 ul:nth-child(3) li:nth-child(2)').text()[6:])

        f_download=response.doc('.summary01 ul:nth-child(3) li:nth-child(3)').text()[6:]

        if not f_download:#for some situations where no f_download
            f_download=int(f_refer)
            f_refer=0
        else:
            f_download=int(f_download)

        try:
            cursor.execute("insert into
paper_cdfd(num,title,author,supervisor,category,type,key_word,year,college,degree,f_download,f_refer,url,abstra
ct) values(%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s)",
[int(num),title,author,supervisor,category,type,key_word,year,college,degree,f_download,f_refer,url,abstract])
            db.commit()
        except Exception as e:
            return (Exception,":",e,':',num)

        return num
```

Figure 16: Partial codes in CNKI crawler

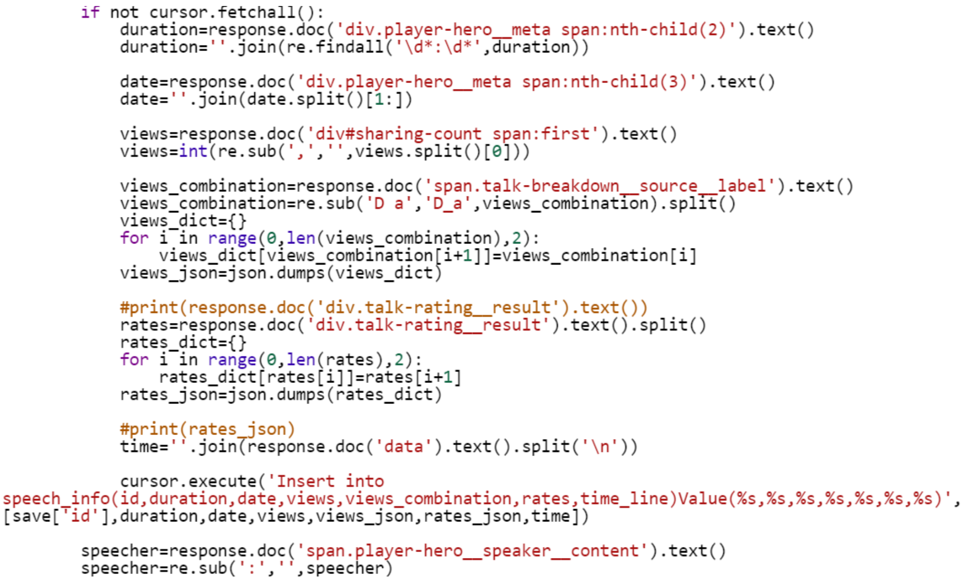
There are numerous important data that worth crawling, such as the author’s degree, abstract, key word, category, times of download. The code capture illustrates how to obtain those data from HTML response. The Python code directly uses JQuery to detect the position of required data rather than apply unreadable regular expression grammars. Moreover, after acquiring target data, they are automatically preserved in the database. The code also displays the corresponding SQL grammar.



```
@every(minutes=24 * 60)
def on_start(self):
    for ab in result:
        for page in range(1,70):
            url='http://www.ted.com/talks?sort=newest&language='+str(ab[0])
            url+='&page='+str(page)
            self.crawl(url, save={'ab':ab},callback=self.index_page,fetch_type='js')

@config(age=10 * 24 * 60 * 60)
def index_page(self, response):
    abbr=response.save['ab']
    print(abbr)
    for each in response.doc('div.media__message a').items():
        r=re.sub('(.*/)|(\?.*)','',each.attr.href)
        cursor.execute('select id from speech_lang where url=%s',r)
        result=cursor.fetchall()
        print(r)
        if result:
            id=int(result[0][0])
        else:
            cursor.execute('select max(id) from speech_lang')
            max=cursor.fetchall()[0][0]
            if max:
                id=int(max)+1
            else:
                id=1
        cursor.execute('Insert into speech_lang(id,url,language) Value(%s,%s,%s)',[id,r,abbr])
        db.commit()

    return abbr
```

Figure 17: Partial codes in TED crawler

Compared with CNKI, the web page structure of TED is more complex. However, for the sake of efficiency, it also needs a fetcher like the CNKI crawler to obtain the number of eligible pages first. The confusion is that one speech in Chinese may not have its corresponding speech in Portuguese. One possible method is to record each Chinese speech title in the database. When searching the Portuguese, only the speeches are crawled. The code above exactly displays the principle of fetcher. Same as CNKI, it automatically stores the result into database so that the crawler server can directly analyse them.



```
        if not cursor.fetchall():
            duration=response.doc('div.player-hero__meta span:nth-child(2)').text()
            duration=''.join(re.findall('\d*:\d*',duration))

            date=response.doc('div.player-hero__meta span:nth-child(3)').text()
            date=''.join(date.split()[1:])

            views=response.doc('div#sharing-count span:first').text()
            views=int(re.sub(',','',views.split()[0]))

            views_combination=response.doc('span.talk-breakdown__source__label').text()
            views_combination=re.sub('D a','D_a',views_combination).split()
            views_dict={}
            for i in range(0,len(views_combination),2):
                views_dict[views_combination[i+1]]=views_combination[i]
            views_json=json.dumps(views_dict)

            #print(response.doc('div.talk-rating__result').text())
            rates=response.doc('div.talk-rating__result').text().split()
            rates_dict={}
            for i in range(0,len(rates),2):
                rates_dict[rates[i]]=rates[i+1]
            rates_json=json.dumps(rates_dict)

            #print(rates_json)
            time=''.join(response.doc('data').text().split('\n'))

            cursor.execute('Insert into
speech_info(id,duration,date,views,views_combination,rates,time_line)Value(%s,%s,%s,%s,%s,%s,%s)',
[save['id'],duration,date,views,views_json,rates_json,time])

        speecher=response.doc('span.player-hero__speaker__content').text()
        speecher=re.sub(':','',speecher)
```

Figure 18: Partial codes in TED crawler

This portion of code is used to acquire some of required data from the HTML response. The required information includes the description, the author, the comment, the date, the rate and foremost transcript. A small problem is that information apart from description and transcript are same even in different languages. Thus, in order to only check them once, it is necessary to identify the related feature. The above code actually does the work. The principle is to record some of the information in the database when it crawls the page in the first time. Then, for each crawling, check the information from the database to judge whether the information has been crawled.

4.2 Scripts Implementation

Most of the programs are written by Python. Some of data analysis and diagram drawing are finished by MATLAB.

The table below lists the overall structure of the scripts concerning this project.

Table 2: Language used for scripts

Language	Python, MATLAB
Package manager	Anaconda3

Packages & Libraries	Built-in	os codec pymysql csv re time urllib
	Third-party	NLTK Numpy
	Self-written	Countlib Cleanlib SAlib BLEU
IDE/tools	PyCharm	

Built-in packages:

- *os*

  It is used to trace the file system so that data can be directly read from text files. Besides, it can also create file and directories. Hence, the crawler middleware uses this package to import data from files and export data to files.
- *codec*

  The package is a file reader. Compared with the embed function *read,* it can intelligently decode and encode the file content into required one. Furthermore, this package uses stream to process the data so that file loading will be very fast.
- *pymysql*

  It is one of the most efficient Python-MySQL database connector. Fast is the best character that it owns.
- *csv*

  This package is utilized to load or create csv file. In this project, csv file is the fundamental format which conserves the data.
- *re*

  This is the most famous regular expression package. Nearly every script imports this package. In crawler, it is used to formatting the result, especially web URL. Besides, it is the essential unit to do data cleaning. Sentence alignment algorithms also make use of regular expression to identify the key punctuation. Even in the evaluation process, it is applied to find the translation result from Google Translation.

- *time*

  It is a less important package which only plays the role of calculating the executing time of algorithms
- *urllib*

  It is a built-in package for web request. In this project, this package is used for obtaining and parsing the response from Google and Baidu translation.

Third-party packages:

- *NLTK*

  NLTK is a very popular Python NLP library. It consists of all kinds of NLP tools, including word segment, sentiment analysis, etc. In terms of this project, two very important libraries are imported from NLTK. One is called *Gale_Church*, the other one is called *Bleu*. *Gale_Church* is a typical length-based sentence alignment algorithm. The off-the-shelf library will be regarded as a contrast against algorithm which is designed by this project. *Bleu* is the other useful off-the-shelf library in NLTK. It is applied as a basic component of translation evaluation function in this project.
- *Numpy*

  It is a well-known scientific computation library which is not only widely used in this project but also in many other computation-based programs.

Self-written packages:

- *Countlib*

  This package is used to count the number of elements in strings. It admits both file and string as input and output.
- *Cleanlib*

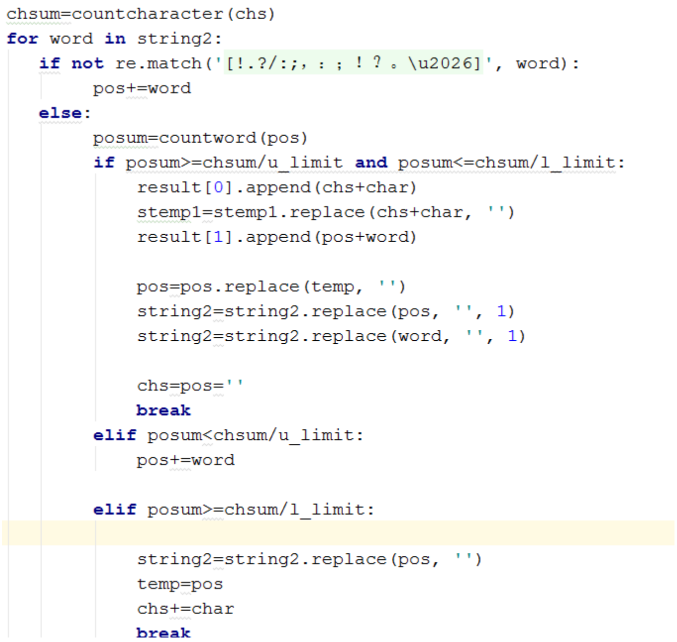
  All of the data cleaning functions including tokenization are comprised inside this package. It also provides different format of input and output. The input can be a file or a string. Compared with the input, the type of output is more various. Cleaning and tokenization are separated, while sentence splitting is also an individual function. There is an ensemble function that does all the pre-processing in a single time. Besides, user can also pick one of them to use.
- *SAlib*

It consists of all the alignment function. Alignment includes paragraph alignment, full-sentence alignment and sub-sentence alignment. The so-called “full-sentence alignment” means align the sentence pair without break complete Chinese sentences. Thus, “sub-sentence alignment” means further align the sentence pair into smaller blocks. As a result, full stop may not be the end of one Chinese sentence, instead comma may be. This package provides an ensemble function which directly executes every step of sentence alignment. Otherwise, user can use them separately.

- *BLEU*

  This package contains the full-automatic translation evaluation functions. It provides direct and indirect translation result and BLEU score individually. User can also obtain single result from Baidu or Google. Input string should be either one sentence pair or a parallel article. Respectively, the outputs are one score and one array of scores.

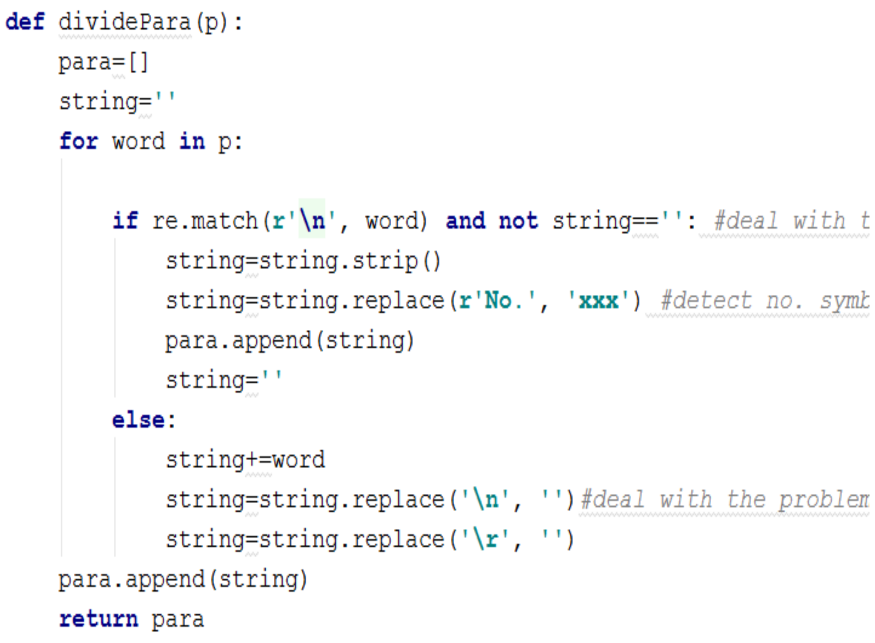
The follow sections demonstrate some of the core scripts.



```
chsum=countcharacter(chs)
for word in string2:
    if not re.match('[!.?/:;，：；！？。\u2026]', word):
        pos+=word
    else:
        posum=countword(pos)
        if posum>=chsum/u_limit and posum<=chsum/l_limit:
            result[0].append(chs+char)
            stemp1=stemp1.replace(chs+char, '')
            result[1].append(pos+word)

            pos=pos.replace(temp, '')
            string2=string2.replace(pos, '', 1)
            string2=string2.replace(word, '', 1)

            chs=pos=''
            break
        elif posum<chsum/u_limit:
            pos+=word

        elif posum>=chsum/l_limit:

            string2=string2.replace(pos, '')
            temp=pos
            chs+=char
            break
```

Figure 19: Kernel codes for sentence alignment

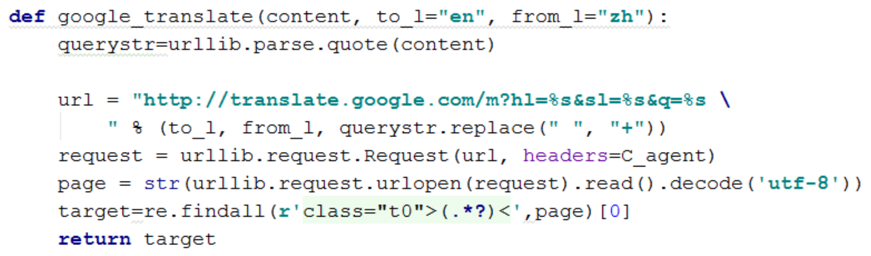
The above codes are the kernel part of the sentence alignment algorithm. Whether align or not depends on the value of the length ratio. If this ratio is out of allowable range, the algorithm will add one more corresponding sentence to the source sentence. The ratio is generated from the length of each sentence. Sentence length is decided by the amount of words in Portuguese or characters in Chinese. However, the ending of sentence is dynamic. In paragraph alignment stage, the ending symbol is line feed, such as “\n”. In full-sentence alignment stage, the ending symbol is full stop, question mark, exclamation mark, etc. In sub-sentence alignment stage, the ending symbol is comma like symbol.



```
def dividePara(p):
    para=[]
    string=''
    for word in p:

        if re.match(r'\n', word) and not string=='': #deal with t
            string=string.strip()
            string=string.replace(r'No.', 'xxx') #detect no. symb
            para.append(string)
            string=''
        else:
            string+=word
            string=string.replace('\n', '')#deal with the problem
            string=string.replace('\r', '')
    para.append(string)
    return para
```

Figure 20: Partial codes for paragraph alignment

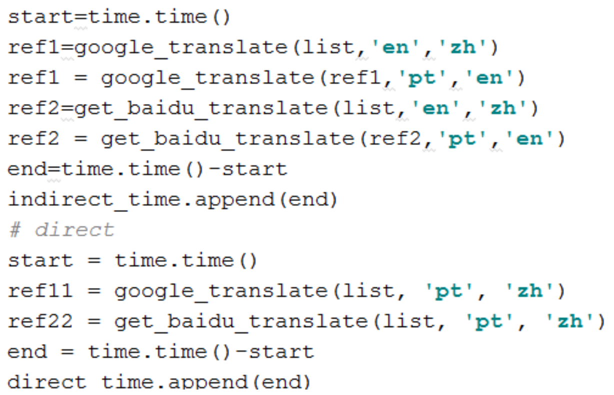
In order to eliminate the computing time, this paragraph alignment function only the one to one case. If the accuracy is more important, ignoring the above code will be a good choice, instead using another iteration of sentence alignment and using “\n” as the ending mark. Note: “\n” is the line feed only in Linux like operation systems. “\n\r” is the line feed of Windows.



```
def google_translate(content, to_l="en", from_l="zh"):
    querystr=urllib.parse.quote(content)

    url = "http://translate.google.com/m?hl=%s&sl=%s&q=%s \
        " % (to_l, from_l, querystr.replace(" ", "+"))
    request = urllib.request.Request(url, headers=C_agent)
    page = str(urllib.request.urlopen(request).read().decode('utf-8'))
    target=re.findall(r'class="t0">(.*?)<',page)[0]
    return target
```

Figure 21: Function to get Google translation result

Rather than register and apply API from Baidu and Google, this script totally utilizes the crawler technology to acquire translation result. The above script displays the function of obtaining result from Google Translation. It requests the website first. Then, parse the response to get required data.



```
start=time.time()
ref1=google_translate(list,'en','zh')
ref1 = google_translate(ref1,'pt','en')
ref2=get_baidu_translate(list,'en','zh')
ref2 = get_baidu_translate(ref2,'pt','en')
end=time.time()-start
indirect_time.append(end)
# direct
start = time.time()
ref11 = google_translate(list, 'pt', 'zh')
ref22 = get_baidu_translate(list, 'pt', 'zh')
end = time.time()-start
direct_time.append(end)
```

Figure 22: Direct and indirect translation

The above codes show the operation of direct translation and indirect translation methods. With the help of pre-defined functions, “google_translate” and “get_baidu_translate”, this task is easy to achieve. Besides, it also records the requesting time into arrays. The time information will help the package user trace the network situation.

4.3 Cloud Platform Implementation

The implementation of this platform includes both the front-end client and back-end server. The client side plays more important role than the server side, since the aim of this platform is to give human translators better workplace. The front-end is written by HTML5, CSS3 and JavaScript, which ensure the compatibility of modern browsers, such as Chrome. Furthermore, the most famous framework called Bootstrap is applied in order to enhance the development efficiency. However, other than the static element, dynamic components are frequently used in the front-end. For example, articles and their sentence pairs are displayed depending on the task number that students request. Compared with the client side GUI, the back-end is designed to achieve some indispensable functions. PHP scripts mostly take responsible for replying the request from the front-end. Meanwhile, Python scripts assist it. Therefore, a cross language service, thrift [32], is applied for the communication of PHP and Python. A well-known PHP Content Management System (CMS) called CodeIgniter [33] is utilized on the server side. By providing mature modules, such as user login, it sharply reduces the development time. Different from local IDE, Cloud9 [34] provides cloud-level web design IDE so that programmer can modify the code everywhere.

The following table illustrates the structure of tools and programming languages used for the platform.

Table 3: Languages and tools used for cloud platform

Development	**Front-end**	**Back-end**
Language	HTML5, CSS3, JavaScript	PHP, Python
Framework	Bootstrap	
CMS	CodeIgniter	
Web server	Apache	
Database	MySQL	
IDE/Tools	PyCharm, Notepad++, Cloud9	
Cross-language	Thrift	

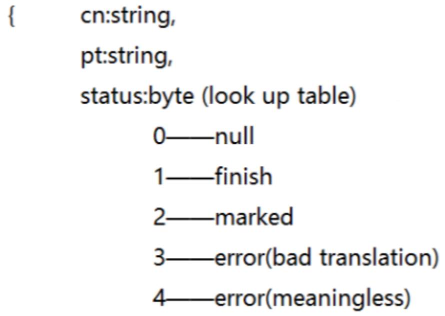
There are several central components inside the platform.

- User workplace
  Translation and sentence alignment modification are two basic tasks of users. This workplace provides both the original monolingual articles (Chinese and Portuguese articles) and sentence-aligned articles. The original articles are used to trace the error reasons if the aligned ones are strange. Besides, users only modify the sentence-aligned articles. Jobs of sentence alignment should not confuse with jobs of translation modification. Thus, while doing alignment, this component enables user to move the sentence or sub-sentence but not modify the content.
- Task assignment component

This component which is written by PHP entirely executes on server side. The function of it is to automatically assign a task to students. It avoids the extra time consumption caused by manually picking tasks. When the system started, it will generate a sorted array, including the id of all articles from the smallest one to the largest one. Each time a student user getting a task, this component will assign the first item of the array to the user and then pop it out. As a result, the process of obtaining a task only costs negligible time. Meanwhile, student user will get the easiest task.

- Task retention component

  It is necessary to retain the task when user leaves the workplace. Otherwise they have to finish the whole task at a single time. Task retention not only records what the user has done, but also mark the errors. After each operation, an ajax request containing json data will be automatically sent to server. The json structure is shown below:



```
{       cn:string,
        pt:string,
        status:byte (look up table)
                0——null
                1——finish
                2——marked
                3——error(bad translation)
                4——error(meaningless)
```

Figure 23: json structure of “bilingual_after”

There are three keys in it. The value of each key is an array. Because array is ordered, it is unnecessary to use another byte to mark the order. Inside status, it is a look-up table. The meaning of each number is displayed clearly.

5 Result and Discussion

In this chapter, the text output, accuracy analysis and cloud platform demonstration will be shown. Detailed discussions are separated in each subsection. Besides, one overall discussion that summarizes all the results is put at the end of this chapter.

5.1 Results

Results will be shown and analysed in four aspects. In terms of original bilingual text, there is only a summary that lists the gross and the related information. Moreover, results from stage two (pre-processing) and stage three (sentence alignment) are in text format. In the third sub-section, charts will be imported to analyse and present the data, especially the translation evaluation result. Last but not least, the GUI of cloud platform is introduced in the end.

5.1.1 Crawled Data

Bilingual text from TED has the best quality. There are totally 2298 speeches published by TED, while 2234 of them have Chinese version and 1804 of them have Portuguese version. After arrangement, there are 1766 Chinese-Portuguese bilingual speeches remained. However, a fatal problem exists in nearly all of those texts. There are meaningless spaces randomly emerge inside the sentence. After tracking, it is unlucky that the phenomenon occurs even in the original translation result. Although space will not influence human reading, it severely damages the computer processing. Even more unfortunately, there is no possible way to eliminate those randomly appeared space by program. Consequently, 105 speeches with higher quality were picked from overall 1766 speeches by hand.

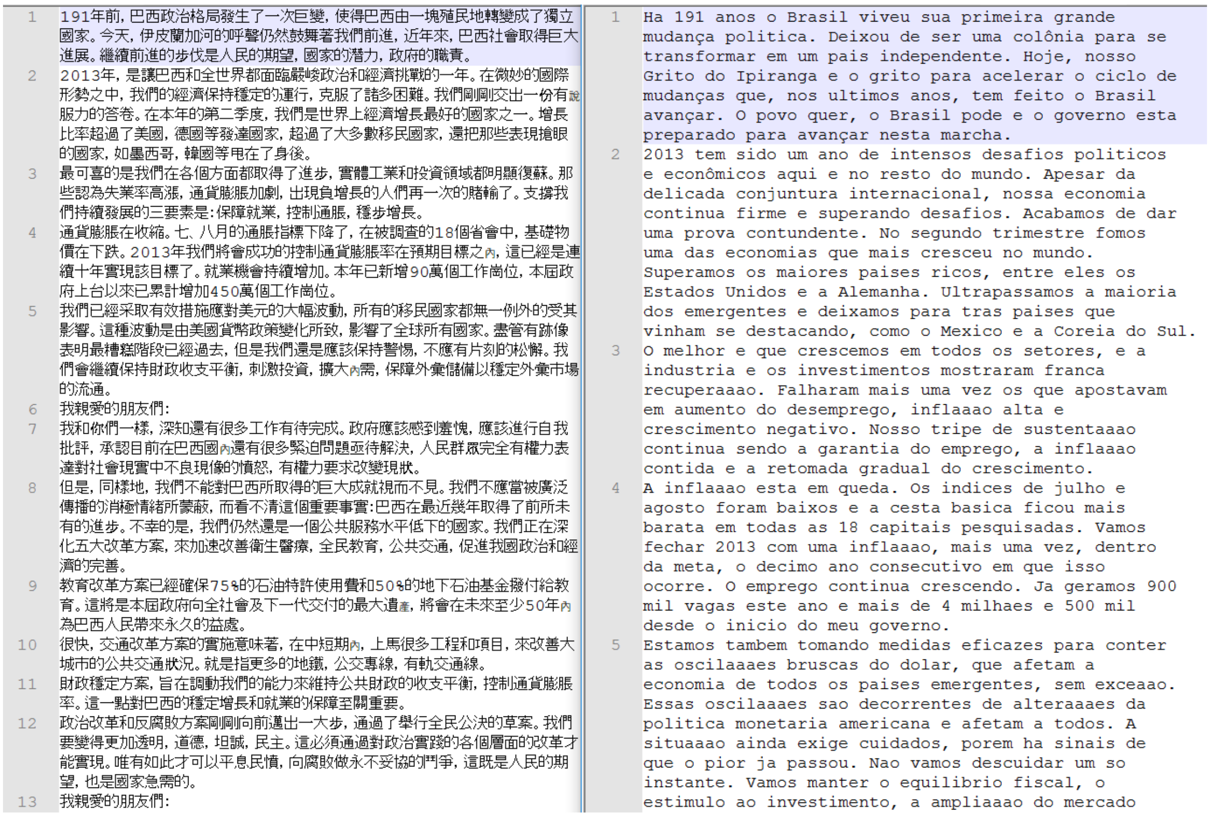
The second part of crawled data which are from Macao websites is quite large but low-quality. The target websites are www.io.gov.mo, www.safp.gov.mo, gb.macaotourism.gov.mo etc. These websites are the only discovered sources that have both Chinese version and Portuguese version. In total, 47030 articles, over 2GB, were stored in the database.

Regarding of CNKI, it is extremely difficult to fully crawl. In the end, only 36650 records have been obtained. All relevant data structures are put in the appendix.

5.1.2 Text Output

Most of the process in data pre-processing and sentence alignment directly acts on texts. Whereas, text does not have gorgeous shape and inspirational effect. This section will only list out some key outputs of both data pre-processing and sentence alignment. The display order will follow the corpus construction procedure.

The output examples are basically generated from a pair of article which is a speech draft from Matt Rosoff, president of Brazil. The original speech is shown below:



1 191年前，巴西政治格局發生了一次巨變，使得巴西由一塊殖民地轉變成了獨立國家。今天，伊皮蘭加河的呼聲仍然鼓舞著我們前進，近年來，巴西社會取得巨大進展。繼續前進的步伐是人民的期望，國家的潛力，政府的職責。

2 2013年，是讓巴西和全世界都面臨嚴峻政治和經濟挑戰的一年。在微妙的國際形勢之中，我們的經濟保持穩定的運行，克服了諸多困難。我們剛剛交出一份有說服力的答卷。在本年的第二季度，我們是世界上經濟增長最好的國家之一。增長比率超過了美國，德國等發達國家，超過了大多數移民國家，還把那些表現搶眼的國家，如墨西哥，韓國等甩在了身後。

3 最可喜的是我們在各個方面都取得了進步，實體工業和投資領域都明顯復蘇。那些認為失業率高漲，通貨膨脹加劇，出現負增長的人們再一次的賭輸了。支撐我們持續發展的三要素是：保障就業，控制通脹，穩步增長。

4 通貨膨脹在收縮。七、八月的通脹指標下降了，在被調查的18個省會中，基礎物價在下跌。2013年我們將會成功的控制通貨膨脹率在預期目標之內，這已經是連續十年實現該目標了。就業機會持續增加。本年已新增90萬個工作崗位，本届政府上台以來已累計增加450萬個工作崗位。

5 我們已經采取有效措施應對美元的大幅波動，所有的移民國家都無一例外的受其影響。這種波動是由美國貨幣政策變化所致，影響了全球所有國家。盡管有跡像表明最糟糕階段已經過去，但是我們還是應該保持警惕，不應有片刻的松懈。我們會繼續保持財政收支平衡，刺激投資，擴大內需，保障外彙儲備以穩定外彙市場的流通。

6 我親愛的朋友們：

7 我和你們一樣，深知還有很多工作有待完成。政府應該感到羞愧，應該進行自我批評，承認目前在巴西國內還有很多緊迫問題亟待解決，人民群衆完全有權力表達對社會現實中不良現像的憤怒，有權力要求改變現狀。

8 但是，同樣地，我們不能對巴西所取得的巨大成就視而不見。我們不應當被廣泛傳播的消極情緒所蒙蔽，而看不清這個重要事實：巴西在最近幾年取得了前所未有的進步。不幸的是，我們仍然還是一個公共服務水平低下的國家。我們正在深化五大改革方案，來加速改善衛生醫療，全民教育，公共交通，促進我國政治和經濟的完善。

9 教育改革方案已經確保75%的石油特許使用費和50%的地下石油基金撥付給教育。這將是本届政府向全社會及下一代交付的最大遺產，將會在未來至少50年內為巴西人民帶來永久的益處。

10 很快，交通改革方案的實施意味著，在中短期內，上馬很多工程和項目，來改善大城市的公共交通狀況。就是指更多的地鐵，公交專線，有軌交通線。

11 財政穩定方案，旨在調動我們的能力來維持公共財政的收支平衡，控制通貨膨脹率。這一點對巴西的穩定增長和就業的保障至關重要。

12 政治改革和反腐敗方案剛剛向前邁出一大步，通過了舉行全民公決的草案。我們要變得更加透明，道德，坦誠，民主。這必須通過對政治實踐的各個層面的改革才能實現。唯有如此才可以平息民憤，向腐敗做永不妥協的鬥爭，這既是人民的期望，也是國家急需的。

13 我親愛的朋友們：

1 Ha 191 anos o Brasil viveu sua primeira grande mudança politica. Deixou de ser uma colônia para se transformar em um pais independente. Hoje, nosso Grito do Ipiranga e o grito para acelerar o ciclo de mudanças que, nos ultimos anos, tem feito o Brasil avançar. O povo quer, o Brasil pode e o governo esta preparado para avançar nesta marcha.

2 2013 tem sido um ano de intensos desafios politicos e econômicos aqui e no resto do mundo. Apesar da delicada conjuntura internacional, nossa economia continua firme e superando desafios. Acabamos de dar uma prova contundente. No segundo trimestre fomos uma das economias que mais cresceu no mundo. Superamos os maiores paises ricos, entre eles os Estados Unidos e a Alemanha. Ultrapassamos a maioria dos emergentes e deixamos para tras paises que vinham se destacando, como o Mexico e a Coreia do Sul.

3 O melhor e que crescemos em todos os setores, e a industria e os investimentos mostraram franca recuperaaao. Falharam mais uma vez os que apostavam em aumento do desemprego, inflaaao alta e crescimento negativo. Nosso tripe de sustentaaao continua sendo a garantia do emprego, a inflaaao contida e a retomada gradual do crescimento.

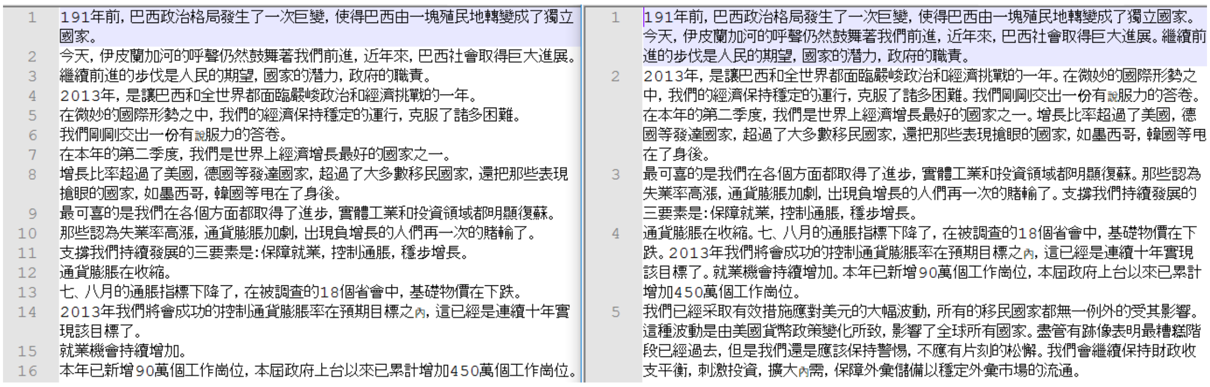
4 A inflaaao esta em queda. Os indices de julho e agosto foram baixos e a cesta basica ficou mais barata em todas as 18 capitais pesquisadas. Vamos fechar 2013 com uma inflaaao, mais uma vez, dentro da meta, o decimo ano consecutivo em que isso ocorre. O emprego continua crescendo. Ja geramos 900 mil vagas este ano e mais de 4 milhaes e 500 mil desde o inicio do meu governo.

5 Estamos tambem tomando medidas eficazes para conter as oscilaaaes bruscas do dolar, que afetam a economia de todos os paises emergentes, sem exceaao. Essas oscilaaaes sao decorrentes de alteraaaes da politica monetaria americana e afetam a todos. A situaaao ainda exige cuidados, porem ha sinais de que o pior ja passou. Nao vamos descuidar um so instante. Vamos manter o equilibrio fiscal, o estimulo ao investimento, a ampliaaao do mercado

Figure 24: Original Chinese (left) and Portuguese (right) articles

Both texts are opened by Notepad++. The left number indicates the paragraph. Due to space difference, there are thirteen paragraphs from Chinese version but only five paragraphs from Portuguese version.

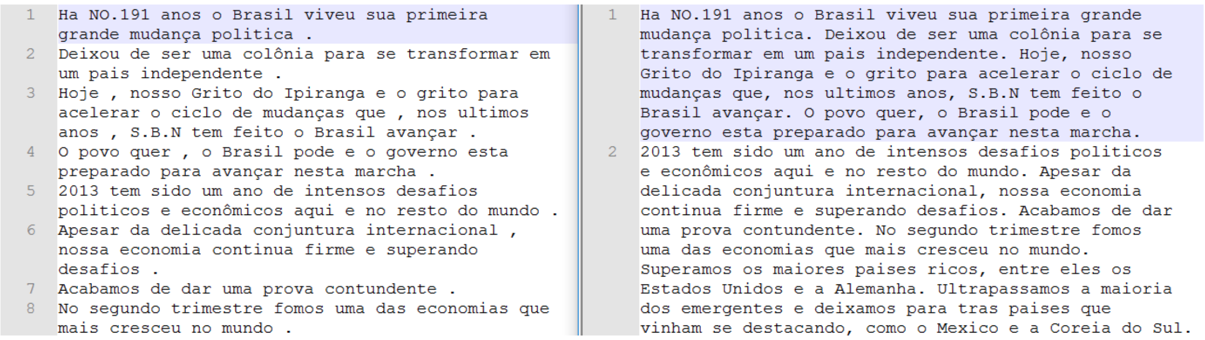
Firstly, it is data pre-processing output.



```
1  191年前，巴西政治格局發生了一次巨變，使得巴西由一塊殖民地轉變成了獨立
   國家。
2  今天，伊皮蘭加河的呼聲仍然鼓舞著我們前進，近年來，巴西社會取得巨大進展。
3  繼續前進的步伐是人民的期望，國家的潛力，政府的職責。
4  2013年，是讓巴西和全世界都面臨嚴峻政治和經濟挑戰的一年。
5  在微妙的國際形勢之中，我們的經濟保持穩定的運行，克服了諸多困難。
6  我們剛剛交出一份有說服力的答卷。
7  在本年的第二季度，我們是世界上經濟增長最好的國家之一。
8  增長比率超過了美國，德國等發達國家，超過了大多數移民國家，還把那些表現
   搶眼的國家，如墨西哥，韓國等甩在了身後。
9  最可喜的是我們在各個方面都取得了進步，實體工業和投資領域都明顯復蘇。
10 那些認為失業率高漲，通貨膨脹加劇，出現負增長的人們再一次的賭輸了。
11 支撐我們持續發展的三要素是：保障就業，控制通脹，穩步增長。
12 通貨膨脹在收縮。
13 七、八月的通脹指標下降了，在被調查的18個省會中，基礎物價在下跌。
14 2013年我們將會成功的控制通貨膨脹率在預期目標之內，這已經是連續十年實
   現該目標了。
15 就業機會持續增加。
16 本年已新增90萬個工作崗位，本屆政府上台以來已累計增加450萬個工作崗位。
```

```
1  191年前，巴西政治格局發生了一次巨變，使得巴西由一塊殖民地轉變成了獨立國家。
   今天，伊皮蘭加河的呼聲仍然鼓舞著我們前進，近年來，巴西社會取得巨大進展。繼續前
   進的步伐是人民的期望，國家的潛力，政府的職責。
2  2013年，是讓巴西和全世界都面臨嚴峻政治和經濟挑戰的一年。在微妙的國際形勢之
   中，我們的經濟保持穩定的運行，克服了諸多困難。我們剛剛交出一份有說服力的答卷。
   在本年的第二季度，我們是世界上經濟增長最好的國家之一。增長比率超過了美國，德
   國等發達國家，超過了大多數移民國家，還把那些表現搶眼的國家，如墨西哥，韓國等甩
   在了身後。
3  最可喜的是我們在各個方面都取得了進步，實體工業和投資領域都明顯復蘇。那些認為
   失業率高漲，通貨膨脹加劇，出現負增長的人們再一次的賭輸了。支撐我們持續發展的
   三要素是：保障就業，控制通脹，穩步增長。
4  通貨膨脹在收縮。七、八月的通脹指標下降了，在被調查的18個省會中，基礎物價在下
   跌。2013年我們將會成功的控制通貨膨脹率在預期目標之內，這已經是連續十年實現
   該目標了。就業機會持續增加。本年已新增90萬個工作崗位，本屆政府上台以來已累計
   增加450萬個工作崗位。
5  我們已經采取有效措施應對美元的大幅波動，所有的移民國家都無一例外的受其影響。
   這種波動是由美國貨幣政策變化所致，影響了全球所有國家。盡管有跡像表明最糟糕階
   段已經過去，但是我們還是應該保持警惕，不應有片刻的松懈。我們會繼續保持財政收
   支平衡，刺激投資，擴大內需，保障外彙儲備以穩定外彙市場的流通。
```

Figure 25: Chinese article after sentence splitting (left) and original article (right)

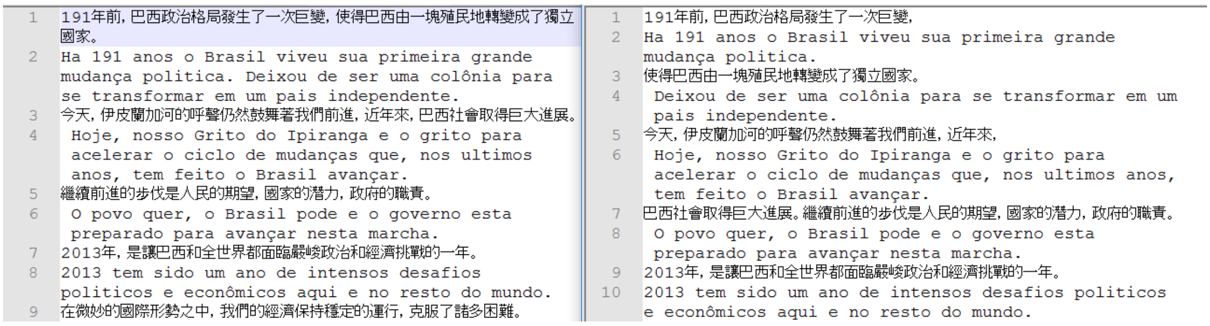


```
1  Ha NO.191 anos o Brasil viveu sua primeira
   grande mudança politica .
2  Deixou de ser uma colônia para se transformar em
   um pais independente .
3  Hoje , nosso Grito do Ipiranga e o grito para
   acelerar o ciclo de mudanças que , nos ultimos
   anos , S.B.N tem feito o Brasil avançar .
4  O povo quer , o Brasil pode e o governo esta
   preparado para avançar nesta marcha .
5  2013 tem sido um ano de intensos desafios
   politicos e econômicos aqui e no resto do mundo .
6  Apesar da delicada conjuntura internacional ,
   nossa economia continua firme e superando
   desafios .
7  Acabamos de dar uma prova contundente .
8  No segundo trimestre fomos uma das economias que
   mais cresceu no mundo .
```

```
1  Ha NO.191 anos o Brasil viveu sua primeira grande
   mudança politica. Deixou de ser uma colônia para se
   transformar em um pais independente. Hoje, nosso
   Grito do Ipiranga e o grito para acelerar o ciclo de
   mudanças que, nos ultimos anos, S.B.N tem feito o
   Brasil avançar. O povo quer, o Brasil pode e o
   governo esta preparado para avançar nesta marcha.
2  2013 tem sido um ano de intensos desafios politicos
   e econômicos aqui e no resto do mundo. Apesar da
   delicada conjuntura internacional, nossa economia
   continua firme e superando desafios. Acabamos de dar
   uma prova contundente. No segundo trimestre fomos
   uma das economias que mais cresceu no mundo.
   Superamos os maiores paises ricos, entre eles os
   Estados Unidos e a Alemanha. Ultrapassamos a maioria
   dos emergentes e deixamos para tras paises que
   vinham se destacando, como o Mexico e a Coreia do Sul.
```

Figure 26: Portuguese article after tokenizing (left) and original article (right)

In terms of Chinese, only sentence splitting is required. Unfortunately, there is no special case in this speech. As a result, it seems that sentences are split only by full stops. The special case will happen if there are quotation marks or brackets. Besides, tokenization is achieved to separate words from punctuations in this case. It is evident that all the punctuation except the abbreviation is surrounded by space in the left side. However, in the right side, punctuations are left-adjacent to one letter. Abbreviations, such as “No.191” on the first line and “S.B.N” above the fourth paragraph, are not separated, which means each of them is identified as one word.



```
1  191年前，巴西政治格局發生了一次巨變，使得巴西由一塊殖民地轉變成了獨立國家。
2  Ha 191 anos o Brasil viveu sua primeira grande mudança politica. Deixou de ser uma colônia para se transformar em um pais independente.
3  今天，伊皮蘭加河的呼聲仍然鼓舞著我們前進，近年來，巴西社會取得巨大進展。
4   Hoje, nosso Grito do Ipiranga e o grito para acelerar o ciclo de mudanças que, nos ultimos anos, tem feito o Brasil avançar.
5  繼續前進的步伐是人民的期望，國家的潛力，政府的職責。
6   O povo quer, o Brasil pode e o governo esta preparado para avançar nesta marcha.
7  2013年，是讓巴西和全世界都面臨嚴峻政治和經濟挑戰的一年。
8  2013 tem sido um ano de intensos desafios politicos e econômicos aqui e no resto do mundo.
9  在微妙的國際形勢之中，我們的經濟保持穩定的運行，克服了諸多困難。
```

```
1  191年前，巴西政治格局發生了一次巨變，
2  Ha 191 anos o Brasil viveu sua primeira grande mudança politica.
3  使得巴西由一塊殖民地轉變成了獨立國家。
4   Deixou de ser uma colônia para se transformar em um pais independente.
5  今天，伊皮蘭加河的呼聲仍然鼓舞著我們前進，近年來，
6   Hoje, nosso Grito do Ipiranga e o grito para acelerar o ciclo de mudanças que, nos ultimos anos, tem feito o Brasil avançar.
7  巴西社會取得巨大進展。繼續前進的步伐是人民的期望，國家的潛力，政府的職責。
8   O povo quer, o Brasil pode e o governo esta preparado para avançar nesta marcha.
9  2013年，是讓巴西和全世界都面臨嚴峻政治和經濟挑戰的一年。
10 2013 tem sido um ano de intensos desafios politicos e econômicos aqui e no resto do mundo.
```

Figure 27: Full-sentence alignment (left) and sub-sentence alignment (right)

The above texts present two levels of alignment. The left one is full-sentence alignment, while the right one is sub-sentence alignment. Compared with the left side, each line on the right may not be stopped by full stop, instead comma. In this way, sentence can be further separated into smaller parts. Nevertheless, the cost is that the aligning accuracy will be reduced. The detail accuracy and the evaluation method will be explained in the next section (5.1.3).

5.1.3 Chart Analysis

Totally three charts are involved in this section, one for the CNKI crawler strategy, one for translation evaluation and the other one for sentence alignment. Furthermore, the accuracy of sentence alignment algorithm which is generated by the translation evaluation metric will be explained in the end of this section.

5.1.3.1 CNKI crawler strategy

As it is mentioned in chapter 3, the CNKI crawler has to take advantage of URL cues. In detail, the id appeared in URL implies the distribution of the valid nodes. Thus, it is possible to prevent searching the wasted nodes with the help of the distribution.

Here is the discrete point diagram:

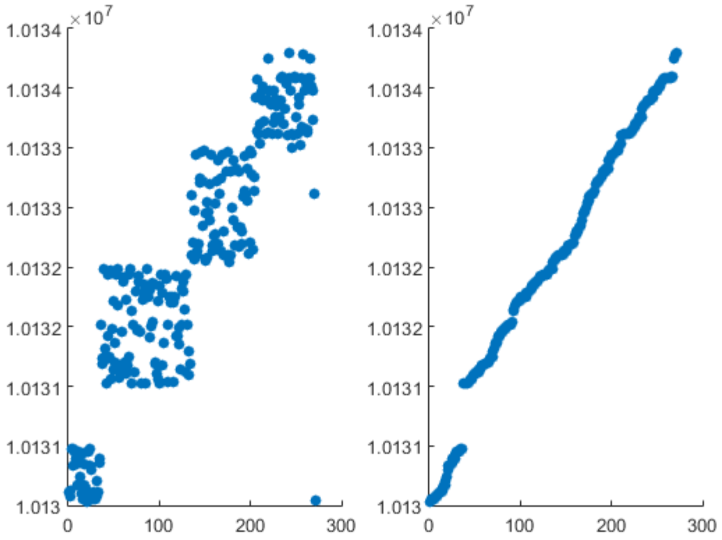


Figure 28: Unsorted scattered points (left) compared with sorted scattered points (right)

The crawler randomly searches 100 nodes out of 10000. The blue point in the diagram shows the id of the valid nodes. The distribution tells that numbers between 1013050 and 1013100 are all invalid, 1013425 and 101350 are scattered. Thus, visiting those URL is unnecessary. Similarly, do such sampling for every range between 0 and 20000. After that, the density region can be gotten. The total crawling points can be sharply reduced, only by searching those URL in the density region.

5.1.3.2 Translation evaluation metric

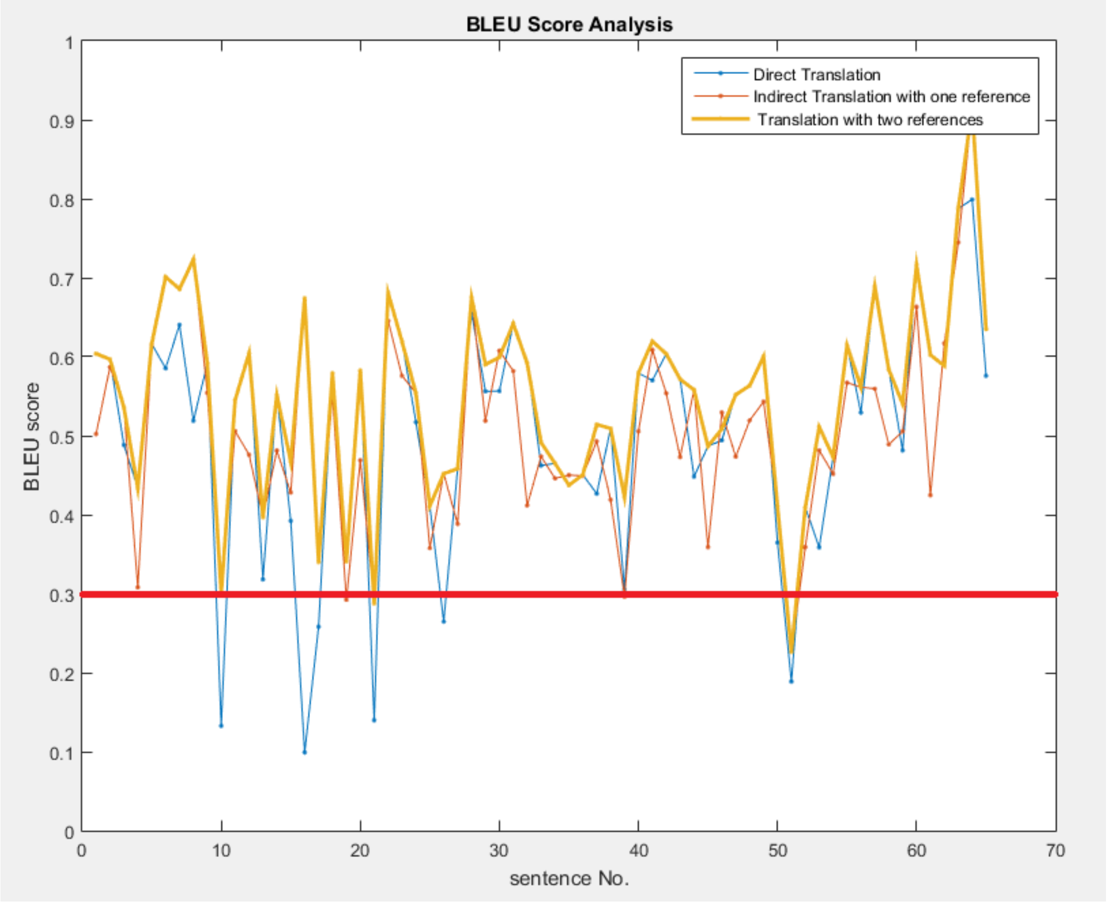


Figure 29: BLEU analysis of the speech

The x-axis indicates the sentence id. For example, one means the first sentence pair. Two means the second sentence pair. The y-axis indicates the BLEU score of the sentence pair. As the legend of this line chart tells, the red line is the BLEU score of indirect translation and the blue line is the BLEU score of direct translation. It is obvious that direct and indirect translation do not have distinguished gap. For some cases the blue line is the upper one and some cases the red line is the upper one. Higher score means the translation is more accurate. Besides, it is known that sentences in the test article are aligned correctly. Therefore, this phenomenon illustrates that either direct translation or indirect translation can be more accurate. No matter which one gets the higher score, it means that the translation is possible to be that score. Hence, the orange line is used to describe the final score.

The bold red line shows the threshold between allowable translation and unacceptable translation. Scores over this line will be treated as acceptable. The 0.3 is a statistical conclusion derived from 10 TED speeches. Only 3 sentences lower than 0.3 out of 64 sentences. Thus, the passing rate of this example is around 0.95.

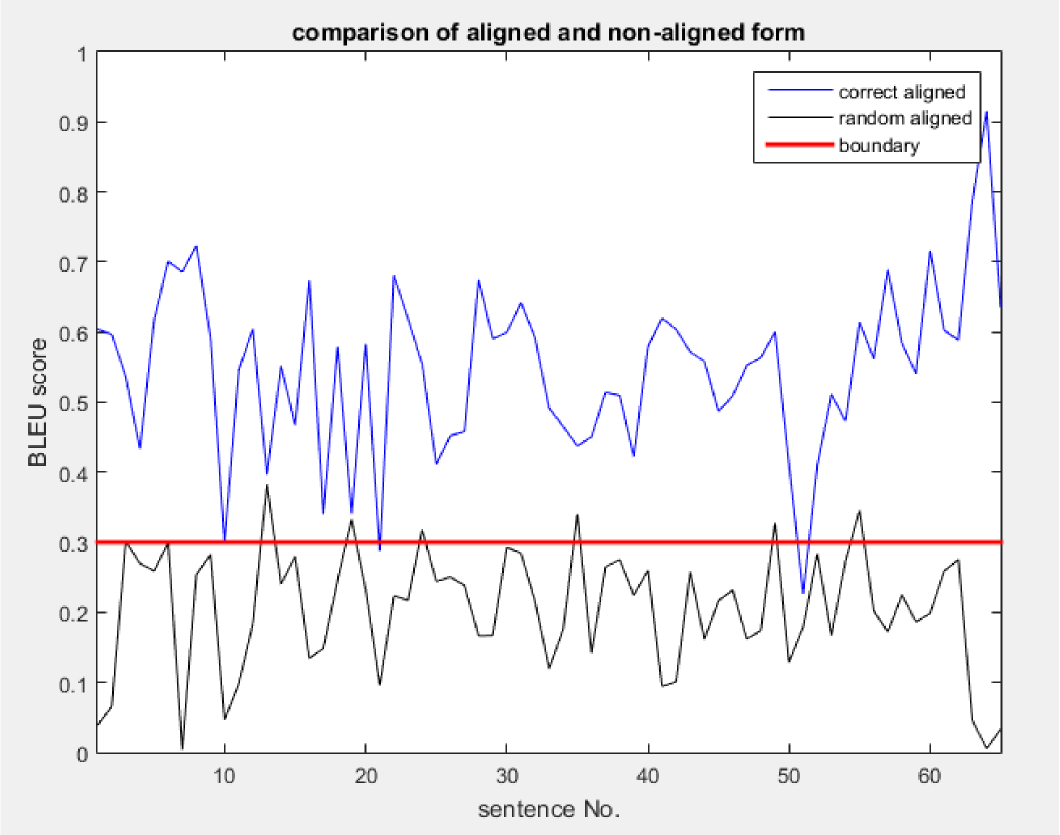


Figure 30: Comparison between correct-aligned and random-aligned article

The above line chart illustrates the comparison between the parallel article and the random-aligned article. The meaning of both x-axis and y-axis are similar to the previous chart. Both blue and black lines come from one identical article. The blue line presents the parallel version, while the black line presents the randomly aligned version. Regarding of the random one, translation pairs are totally meaningless. It is obvious that scores are sparse in the middle of two lines. The bold line with red colour is the threshold that is used to judge the translation quality. If the score is bigger than 0.3, the corresponding translation will be accepted. Otherwise, it will be rejected. The diagram proves that 0.3 is a reasonable boundary. Furthermore, the translation evaluation metric is also testified that it sensitive enough to evaluate the translation quality.

Compared with original BLEU, the new method provided by this project is full-automatic, since it replaces the human translation reference by machine translation reference from

Google and Baidu. As is known to all, the quality of machine translation even Google is far lower than human translation. Consequently, this method can only be applied for roughly estimating the fidelity of the translation result. Fidelity in language translation indicates whether each word in one sentence is translated to the proper word in another sentence or not. In some extent, if the fidelity is higher, even though the overall translation quality is bad, it is still possible to say that two of those are translated to each other. Hence, the metric will be a good tool to evaluate the sentence alignment quality. Meanwhile, it is an auxiliary tool for translation quality evaluation.

5.1.3.3 Sentence alignment effectiveness

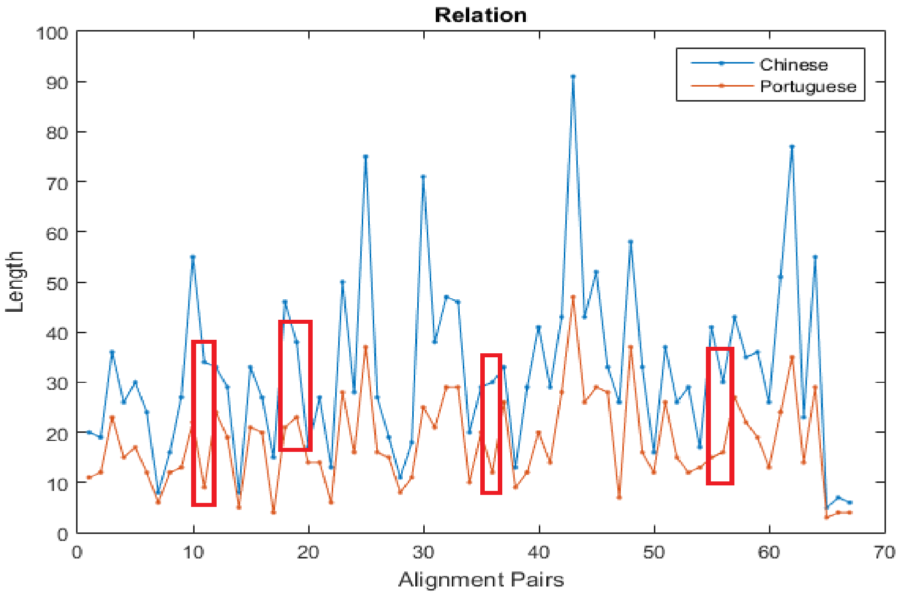


Figure 31: Length-based alignment analysis

This line chart describes the sentence length relationship between each alignment pairs of the example article. The x-axis indicates the number of parallel sentence pairs. For example, x is one, which means x is the first alignment pair, two means the second alignment pair. The parallel sentence pair means they have already been correctly aligned. The y-axis indicates the length of single language of an alignment pair. The blue line shows the Chinese length while the orange line shows the Portuguese length.

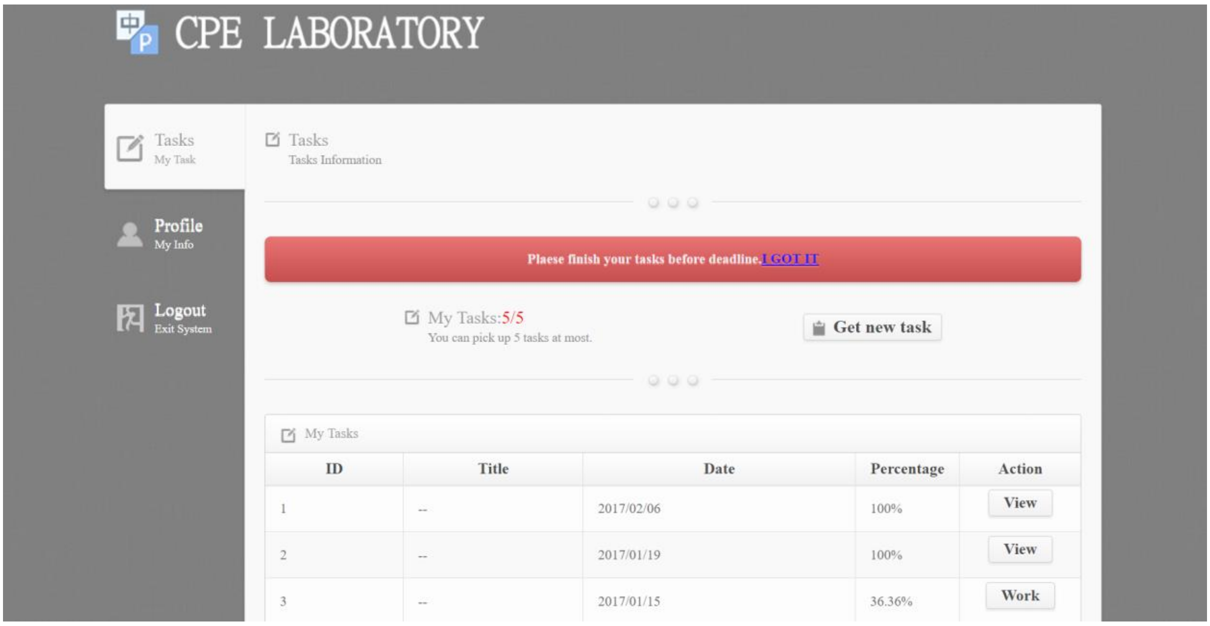
It is evident that for most of the alignment pair the length of each language is in similar correlation. If one becomes bigger, the other one becomes bigger as well. If one becomes lower, the other one will also become lower. This phenomenon implies that length can be a reasonable cue for sentence alignment. What the sentence alignment algorithm needs to do is to find the correlation for each alignment.

However, there are four places which do not obey the correlation shown in red boxes. Such situation indicates that there is an upper accuracy limit, since strange correlation may happen in some specially cases. Luckily, only 4 cases out of 65 alignment pairs occur in this example. The upper accuracy is 0.94 in this case. It also reminds us that the upper limit is so high that will not badly effects the algorithm accuracy.

In consequence, the average passing rate of the algorithm evaluated by the full-automatic BLEU (the metric designed in this project) is 84.5% from 105 TED speeches. It is worth mentioning that such passing rate is not the same meaning as the accuracy. The passing rate which is generated from the translation evaluation algorithm is not a rigorous assessment metric but a rough one. Strictly speaking, it can be regarded as a highly relevant reference. Because passing rate may be depressed due to bad translation quality, the real accuracy, if it is possible to obtain, is possible to be higher the above percentage. To evaluate the real accuracy, it must consume much human resource, which will only be attempted in the future.

5.1.4 Cloud Platform Demonstration

The platform is on the local network of IPM with the domain of catcms.ipm.edu.mo/CorpusManage. Task obtainment UI and workplace UI are two essential UI of the platform.



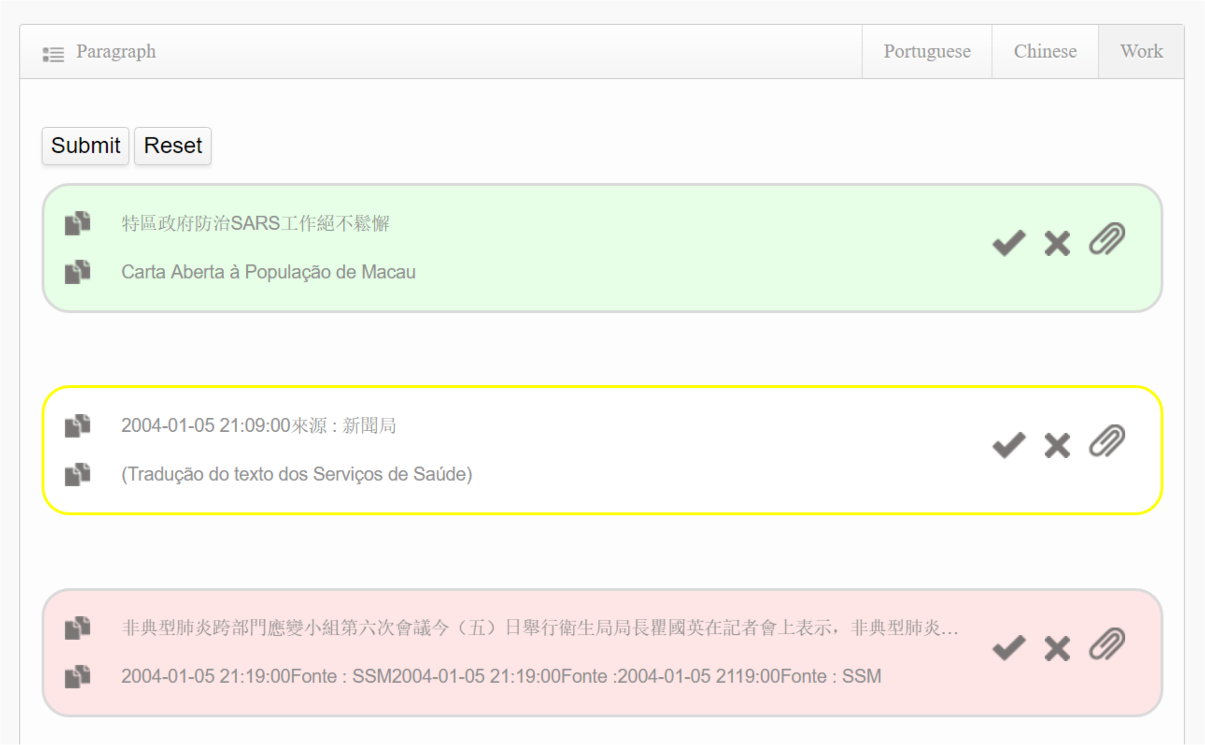
Plaese finish your tasks before deadline.I GOT IT

My Tasks:5/5
You can pick up 5 tasks at most.

ID	Title	Date	Percentage	Action
1	--	2017/02/06	100%	View
2	--	2017/01/19	100%	View
3	--	2017/01/15	36.36%	Work

Figure 32: UI of task obtainment

After logging in by student accounts, the interface will be shown. The student user can get new tasks by clicking the button “get new task” if the uncompleted tasks are less than five. The system will automatically assign a new task for the user. Under the button, it is “My tasks” box which contains all the tasks that the user gained. The user can view tasks that have already been finished or continue working on those unfinished tasks. In this case, the title is empty because most of crawled article does not have a proper title.



特區政府防治SARS工作絕不鬆懈

Carta Aberta à População de Macau

2004-01-05 21:09:00來源 : 新聞局

(Tradução do texto dos Serviços de Saúde)

非典型肺炎跨部門應變小組第六次會議今（五）日舉行衛生局局長瞿國英在記者會上表示，非典型肺炎...

2004-01-05 21:19:00Fonte : SSM2004-01-05 21:19:00Fonte :2004-01-05 2119:00Fonte : SSM

Figure 33: UI of workplace

There are three views of workplace, two monolingual views and one bilingual view. This screen capture displays the bilingual view of the workplace. In the bilingual view, sentences of different language have already been aligned by the sentence alignment algorithm. Users can directly modify the translation result or mark them. There are mainly three kinds of marks shown on the right. The tick, which means finish, will highlight the textbox in green background. The cross, which means error, will highlight the textbox in red background. The clip, which means uncertainty, will highlight the textbox in yellow border. For each operation, the client side will send an ajax request to the server to record the current progress of the user. Thus, users do not need to worry about missing data.

5.2 Overall Discussion

Since most of the outcomes are closed related to the objectives that mentioned in the first chapter, the discussion will also focus on those objectives.

1. *Collecting monolingual or bilingual corpus from the Internet*

   The crawling results from TED and websites from Macao government are bilingual corpora. The crawling result from CNKI is Chinese monolingual corpus. The total corpus size is over 2 GB with at least forty thousand articles. Besides the data, the crawler server is also functional. Compared with individual crawlers, it integrates miscellaneous crawlers together with cohesive crawling schedules. As long as the server runs, it will automatically execute crawlers following the advance setting schedule. In addition, adding new crawlers to the server will be an easy job. Because data crawled from the Internet will be directly stored into database, there is no need to do second-arrangement. Therefore, the flexibility and the expansibility are the primary advantages of this crawler server. Hence, this objective has been fulfilled.

2. *Fulfilling robust data pre-processing programs*

   The data pre-processing program can successfully clean the meaningless items, eliminate duplicated sentences and tokenize text contents. The tokenization program focus on alphabetic languages. Besides the common case, different abbreviation rule can be easily added to the program. Thus, the tokenizer possesses the ability to expand on other alphabetic languages. As a result, it achieves the goal of pre-processing.

3. *Achieving self-designed sentence alignment algorithm*

   Different from all the methods emerged in published papers, the algorithm is absolutely innovative. As results generated from this algorithm have gotten over 80% passing rate, it is confident to prove that this algorithm is workable and effective enough. Furthermore, this algorithm is compatible with different language pairs. Compared with lexicon-based algorithm, it does not need any dictionaries. Considering the outcome, this objective is also fulfilled.

4. *Developing evaluation metrics for both sentence alignment and translation*

   The evaluation metric based on BLEU performs well on the fidelity evaluation. Hence, the metric will be a good tool to evaluate the sentence alignment quality. Meanwhile, it is an auxiliary tool for translation quality evaluation. With the help of this score, human translators will know the rough quality of each translation so that increases their working efficiency. Therefore, the challenge of this objective has successfully been overcome.

5. *Constructing a cloud platform for manually modification*

The cloud platform together with the database has already been set up on the MPI server. Users who have the account can directly log in and start working. Sentence-aligned workplace is the most remarkable design. In order to reduce the workload, the possible alignments which are generated from sentence alignment algorithm will be given as placeholder. For the sake of efficiency, the system tries to provide anything useful to user. Therefore, this objective is also accomplished well.

6 Conclusion and Further Works

This project accomplished an innovative bilingual corpus construction process, from bilingual corpus collection to human modification platform. Along those procedures, it carried out some of the core challenges that corpus construction needed, including data cleaning, sentence alignment and translation evaluation. The sentence alignment algorithm and translation evaluation methods in this project are firstly designed and integrated. Compared with other mature algorithms, this project can be considered as an innovative and reasonable attempt, and it has solved the problems at the initial stage of Chinese-Portuguese bilingual corpus construction.

Bilingual corpus construction in last generation is entirely collected and integrated only by human. Although it has high quality, the whole process consumes massive time. Comparatively, this project belongs to the second generation of corpus construction, which integrates the language knowledge with the computer technology. As a result, it greatly enhances the efficiency meanwhile keeps the quality in a relatively high standard. Cooperating with human, the computer program focuses on completing tedious work and providing cohesive workplaces to human experts. In addition, programs, such as the translation evaluation tool, play the role as assistants which can help human make decisions.

To sum up, on the one hand, from text collection to corpus generation, what the project has done not only replaces most of the artificial work, but also optimizes the final result quality. On the other hand, this project involves in both research and engineering. The academic part enables the project to be expanded on larger scope and gives others inspirational ideas. Meanwhile the industrial part provides a portal that shrinks the distance between theory and practice. With the assist of the transition, the bilingual corpus construction is easily to initiate.

Further works

It is indispensable that some parts of the process need to be optimized in the future. In terms of the first step, the crawler target is fully selected by human. Since the Chinese-Portuguese corpus is extremely scarce, one or two websites will not obviously solve the problem. Thus, an automatic corpus detective may be possible to augment the corpus collection speed. Based on the techniques of search engine, intelligent crawler can be implemented. Furthermore, this project only implemented a document level redundancy eliminator. However, in the real case, document level is far from enough. Although corpus level bilingual duplication detector contains lots of challenges, such as memory limitation when deal with tremendous texts, it is still possible to build one component that can perform the similar function.

References

[1] Wikipedia Corpus Linguistics https://en.wikipedia.org/wiki/Corpus_linguistics, [April 2017]

[2] Xin Shan Ding. The development and current research situation of corpus linguistics. Modern Linguistics. 1998, the first issue.

[3] Koehn P. Europarl: A parallel corpus for statistical machine translation[C] MT summit. 2005, 5: 79-86.

[4] Wikipedia Statistical Machine Translation https://en.wikipedia.org/wiki/Statistical_machine_translation#Sentence_alignment, [April 2017]

[5] Wikipedia Machine translation https://en.wikipedia.org/wiki/Machine_translation, [April 2017]

[6] Kennedy G. An introduction to corpus linguistics [M]. Routledge, 2014.

[7] Wikipedia Natural Language Processing https://en.wikipedia.org/wiki/Natural_language_processing, [April 2017]

[8] Sinclair J. Corpus, concordance, collocation [M]. Oxford University Press, 1991.

[10] Yiping Li, Chiman Pun & Fei Wu (1999), Portuguese-Chinese Machine Translation in Macao

[11] Koehn P. Europarl: A parallel corpus for statistical machine translation[C] MT summit. 2005, 5: 79-86.

[12] Philipp Koehn (2009). Statistical Machine Translation. Cambridge University Press. p. 27. ISBN 0521874157.

[13] Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation[C] Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. Association for Computational Linguistics, 2007: 177-180.

[14] Gale, William.A. & Kenneth W.Church(1993), A Program for Aligning Sentences in Bilingual Corpora, in Using Large Corpora. The MIT Press

[15] Collins M. Statistical machine translation: IBM models 1 and 2[J]. Columbia Columbia Univ, 2011.

[16] NLTK Natural Language Toolkit http://www.nltk.org/, [April 2017]

[17] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C] Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 311-318.

[18] Nießen S, Och F J, Leusch G, et al. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research[C] LREC. 2000.

[19] Levenshtein, Vladimir I. (February 1966). "Binary codes capable of correcting deletions, insertions, and reversals". Soviet Physics Doklady. 10 (8): 707–710.

[20] Resnik P. Parallel strands: A preliminary investigation into mining the web for bilingual text[C] Conference of the Association for Machine Translation in the Americas. Springer Berlin Heidelberg, 1998: 72-82.

[21] Brown P F, Lai J C, Mercer R L. Aligning sentences in parallel corpora[C] Proceedings of the 29th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1991: 169-176.

[22] Warwick S, Russell G. Bilingual concordancing and bilingual lexicography[C] EURALEX 4th International Congress. 1990.

[23] Simard M, Foster G F, Isabelle P. Using cognates to align sentences in bilingual corpora[C] Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2. IBM Press, 1993: 1071-1082.

[24] Sun L, Du L, Sun Y, et al. Sentence Alignment of English-Chinese Complex Bilingual Corpora[C] Proceeding of the workshop MAL. 1999, 99: 135-139.

[25] Ma X, Liberman M. Bits: A method for bilingual text search over the web[C] Machine Translation Summit VII. 1999: 538-542.

[26] Yang C C, Li K W. Building parallel corpora by automatic title alignment using length-based and text-based approaches[J]. Information processing & management, 2004, 40(6): 939-955.

[27] Fei Wang. The design and implementation of bilingual parallel alignment platform. Nanjing University of Science and Technology. 2004

[28] Santorini B. Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision) [J]. 1990.

[29] E.H. Hovy. 1999. Toward finely differentiated evaluation metrics for machine translation. In Proceedings of the Eagles Workshop on Standards and Evaluation, Pisa, Italy.

[30] Baidu Translation API Document http://api.fanyi.baidu.com/api/trans/product/index, [April 2017]

[31] Google Cloud Platform Translation API Price https://cloud.google.com/translate/pricing, [April 2017]

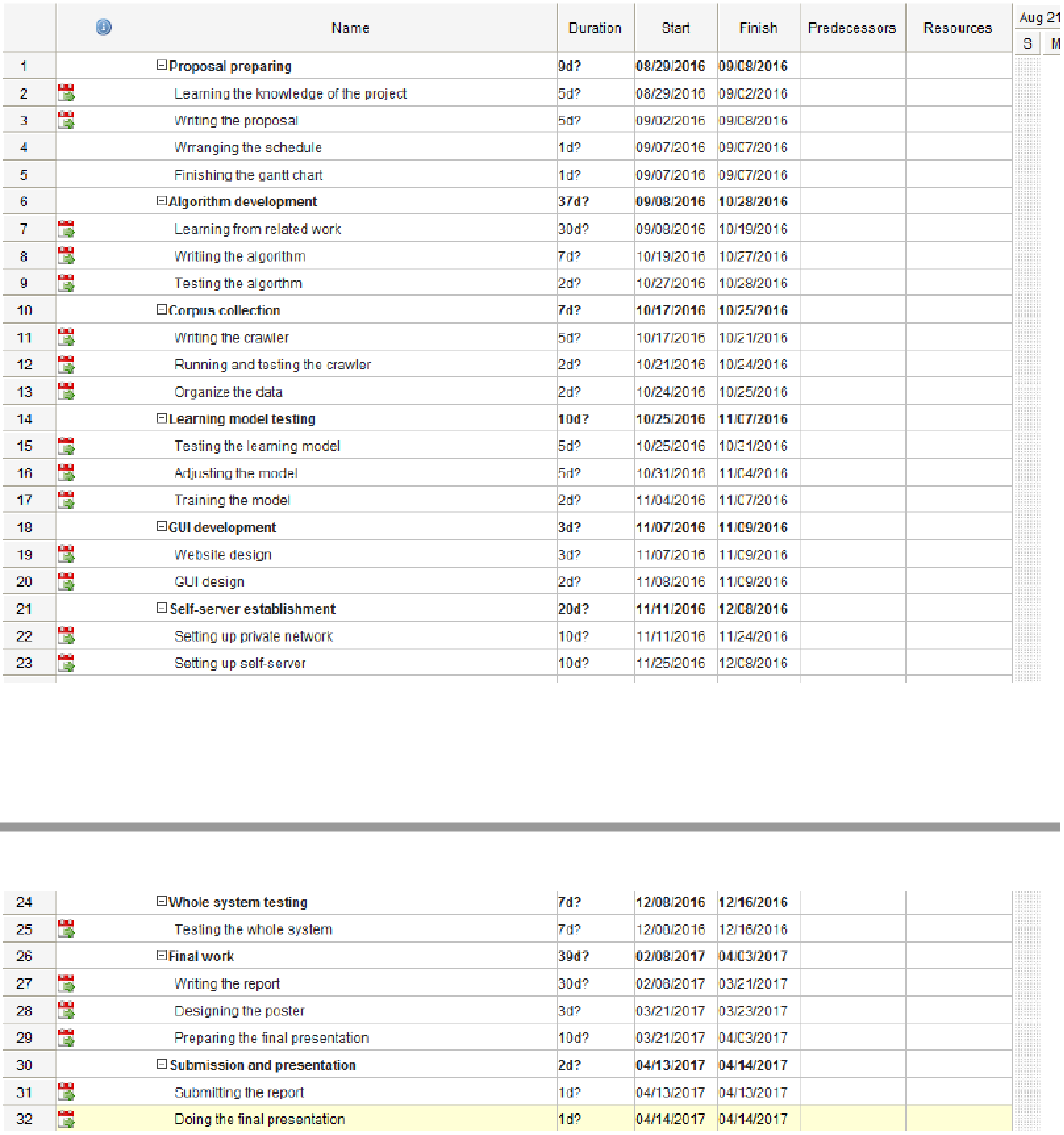
[32] Apache Thrift Official Website http://thrift.apache.org/, [April 2017]

[33] CodeIgniter Official Website http://codeigniter.org.cn/, [April 2017]

[34] Cloud9 Official Website https://c9.io/, [April 2017]

7 Appendix

7.1 Gantt Chart



	ⓘ	Name	Duration	Start	Finish	Predecessors	Resources	Aug 21 S M
1		**⊟Proposal preparing**	**9d?**	**08/29/2016**	**09/08/2016**			
2	📅	Learning the knowledge of the project	5d?	08/29/2016	09/02/2016			
3	📅	Writing the proposal	5d?	09/02/2016	09/08/2016			
4		Wrranging the schedule	1d?	09/07/2016	09/07/2016			
5		Finishing the gantt chart	1d?	09/07/2016	09/07/2016			
6		**⊟Algorithm development**	**37d?**	**09/08/2016**	**10/28/2016**			
7	📅	Learning from related work	30d?	09/08/2016	10/19/2016			
8	📅	Writing the algorithm	7d?	10/19/2016	10/27/2016			
9	📅	Testing the algorthm	2d?	10/27/2016	10/28/2016			
10		**⊟Corpus collection**	**7d?**	**10/17/2016**	**10/25/2016**			
11	📅	Writing the crawler	5d?	10/17/2016	10/21/2016			
12	📅	Running and testing the crawler	2d?	10/21/2016	10/24/2016			
13	📅	Organize the data	2d?	10/24/2016	10/25/2016			
14		**⊟Learning model testing**	**10d?**	**10/25/2016**	**11/07/2016**			
15	📅	Testing the learning model	5d?	10/25/2016	10/31/2016			
16	📅	Adjusting the model	5d?	10/31/2016	11/04/2016			
17	📅	Training the model	2d?	11/04/2016	11/07/2016			
18		**⊟GUI development**	**3d?**	**11/07/2016**	**11/09/2016**			
19	📅	Website design	3d?	11/07/2016	11/09/2016			
20	📅	GUI design	2d?	11/08/2016	11/09/2016			
21		**⊟Self-server establishment**	**20d?**	**11/11/2016**	**12/08/2016**			
22	📅	Setting up private network	10d?	11/11/2016	11/24/2016			
23	📅	Setting up self-server	10d?	11/25/2016	12/08/2016			
24		**⊟Whole system testing**	**7d?**	**12/08/2016**	**12/16/2016**			
25	📅	Testing the whole system	7d?	12/08/2016	12/16/2016			
26		**⊟Final work**	**39d?**	**02/08/2017**	**04/03/2017**			
27	📅	Writing the report	30d?	02/08/2017	03/21/2017			
28	📅	Designing the poster	3d?	03/21/2017	03/23/2017			
29	📅	Preparing the final presentation	10d?	03/21/2017	04/03/2017			
30		**⊟Submission and presentation**	**2d?**	**04/13/2017**	**04/14/2017**			
31	📅	Submitting the report	1d?	04/13/2017	04/13/2017			
32	📅	Doing the final presentation	1d?	04/14/2017	04/14/2017			

Figure 34: the project Gantt chart

7.2 Crawler Data Structure

Table 4: Data structure of bilingual text from Macao websites

Field	Comment	Data Type	PK/FK	Nullable
Num	Unique document id	Int(11)	PK	No
Chinese	Chinese version of the document	Longtext		No
Portuguese	Portuguese version of the document	Longtext		No
Bilingual	Sentence alignment result	Longtext		No
Cn_len	The sentence length of Chinese	Int(8)		Yes
Pt_len	The sentence length of Portuguese	Int(8)		Yes
Len_ratio	Quotient of Cn_len over Pt-len	Float(5,2)		Yes
BLEU	BLEU score of bilingual	Float(3,2)		Yes

Table 5: Data structure of "speech_info" table

Field	Comment	Data Type	PK/FK	Nullable
id	Unique speech id	Int(10)	PK/FK	No
Duration	The duration of the speech	Varchar(6)		Yes
Data	The Speech date	Varchar(6)		Yes
Topic	Topics that the speech involves	Tinytext		Yes
Views	The number of view	Int(12)		Yes
View_combination	Platforms that people use to view the speech	Json		Yes
Rates	The rate of the speech with some give metric	Json		Yes
Comments	The speech comments	Json		Yes
Time_line	Time inside the speech corresponding to transcript each paragraph	Varchar(255)		Yes

Table 6: Data structure of "speech_lang" table

Field	Comment	Data Type	PK/FK	Nullable
id	Speech id	Int(10)	FK	Yes
URL	part of URL unique to the others	Varchar(100)	PK	No
Title	The Speech title	Varchar(150)		Yes
Speecher	The speecher	Varchar(100)		Yes
Description	Brief description	Text		Yes
Language	Language abbreviation	Varchar(10)		No
Transcript	Subtitle of the speech	Text		Yes

8 Reflection

The worthy experience I got from this project is the self-learning skill and the exploration ability. This project has suffered a tortuous confirmation process during the first few months because of the complexity of the research work. In the beginning, I decided to do word classification only with a self-designed weight factor. However, when I worked in the translation laboratory, I got the second research direction, called sentence alignment, which finally became the current topic. Although word classification and sentence alignment are both related to corpus, they are separated field in NLP. As a result, I eliminated one, word classification, from the whole topic. Nevertheless, compared with engineering, research work is more rigorous with limited time and resource. Besides, as the work of translation laboratory is concerning bilingual corpus construction, the final project combines them and involves in both industrial level and academic level.

Corpus construction is not a work of individual but an organization. However, the framework can be initially designed and achieved, which is what this project have done. During my project, I have studied corpus linguistic, sentence alignment and translation evaluation and read lots of relevant papers. Generally speaking, it opens up the gate of computational linguistic to me and let me touch the most attractive part.