Chinese-English Parallel Corpus Construction and its Application

CHANG Baobao
Institute of Computational Linguistics
Peking University
Beijing, P.R.China
chbb@pku.edu.cn

Abstract

Chinese-English parallel corpora are key resources for Chinese-English cross-language information processing, Chinese-English bilingual lexicography, Chinese-English language research and teaching. But so far large-scale Chinese-English corpus is still unavailable yet, given the difficulties and the intensive labours required. In this paper, our work towards building a large-scale Chinese-English parallel corpus is presented. We elaborate on the collection, annotation and mark-up of the parallel Chinese-English texts and the workflow that we used to construct the corpus. In addition, we also present our work toward building tools for constructing and using the corpus easily for different purposes. Among these tools, a parallel concordance tool developed by us is examined in detail. Several applications of the corpus being conducted are also introduced briefly in the paper.

1 Introduction

Corpora are no doubt key resources for language information processing, language research and teaching and lexicography. Many large-scale monolingual corpora have been available in the world, especially with English corpora. Efforts to build Chinese corpora started around 1990s in Mainland China. Substantial progress has been made so far. There are several large-scale Chinese corpora available now, such as the People's Daily corpus developed by Peking University. However, Compared with the big progress in building monolingual corpus, there are not so many practical parallel corpora available, especially parallel corpora with Chinese involved, which hinders their use in cross-language information processing, language teaching and research and bilingual dictionary compilation. In order to give substantial support to research in related fields, we started compiling a Chinese-English parallel corpus in 2001. So far over 10 million character of Chinese texts of different types and corresponding English texts has been collected and included into the corpus. The corpus has been automatically aligned and verified manually at sentence level and the Chinese part also has been automatically segmented and POS tagged. Tentative researches using this corpus in Machine Aided Translation, bilingual dictionary compilation and other related field has been also conducted. We will present our work towards building and using the Chinese-English corpora in this paper.

2 The construction of the Chinese-English corpus

2.1 The collection of the parallel texts

There is a wide literature on corpus design principles; one of them is that a general corpus should be made balanced. Which means a corpus shall contain texts of different domain and different genres in reasonable proportions; the corpus thus can be a reasonable reflection of the language use. However, when we decided to construct the Chinese-English corpus, we found it's not easy to construct a perfectly balanced Chinese-English corpus. That's because there are not so many electronic Chinese-English bilingual texts available. So we decide to collect Chinese-English bilingual texts as many as we can, as

long as the texts are of good quality. For the same reason, we decided not to use any sampling techniques; the full texts are included in the corpus.

So far, over 10 million character of Chinese texts and their corresponding English texts has been collected and included into the corpus. Most of the texts are collected from Internet and cover a variety of domains, such as newspaper news, technical articles, literature, movie transcription and so on. Text noises, i.e. irrelevant HTML tags, figures, tables etc., are removed from the texts before the texts entered into the corpus.

2.2 Annotation of the parallel corpus

The corpus could only be useful after they are annotated. For example, parallel corpus could be used as a Translation Memory only after they are aligned at different level. For a parallel corpus, the most important annotation will be alignment, especially sentence alignment, which will be a minimal requirement for a parallel corpus. In addition, other annotations applicable to monolingual corpus are also applicable to parallel corpus, such as POS tagging and parsing. But annotation of large-scale corpus is labour intensive, which could be done only with the help of automatic tools. Considering of the reliability of the corpus tools and possible use of the corpus, we decided to carry out the following three types of annotation:

1) Global textual attributes. Global textual attributes are attributes applied to every full text in the corpus. They are features to specify the domain of the texts, whether a text is written or spoken, the author of a text, the translator of a text, the time period when a text was authored, the title of a text and so on. The global textual attributes will facilitate special research based on the corpus, for example, language researchers might be interested only with texts belonging to a particular domain, and they can easily extract all texts belonging to that domain.

2) Monolingual textual structural annotation. Monolingual textual structural annotation deals with text unit of different levels. At present, boundaries of paragraph, sentence, and word have been annotated in the corpus. Annotation of word boundaries of Chinese texts is also known as word segmentation. In addition, basic linguistic information such as part of speech types is also tagged for texts of both languages.

3) Parallel alignment annotation. Parallel alignment annotation establishes the correspondence between the language units of the original texts and their translations. So far, the corpus is aligned only at the sentence level. Word alignment of the corpus seems still unpractical for the massive labour required and lacking of reliable tools.

2.3 The mark-up of the parallel corpus

To make the corpus application-independent and easier to exchange via the Internet, all the texts in the corpus shall be encoded uniformly. For this reason, an XML-based framework has been designed and applied to the corpus. According to this framework, all the Chinese texts and English texts are encoded separately, each text, no matter what language it is, is composed of a text head and a text body. All the global textual attributes are put into the text head; the monolingual structural tags, linguistic information tags and the text itself are put within the text body. Alignments are established via the id number recorded in the text body of the both languages. Figure-1 shows a fragment of the corpus.

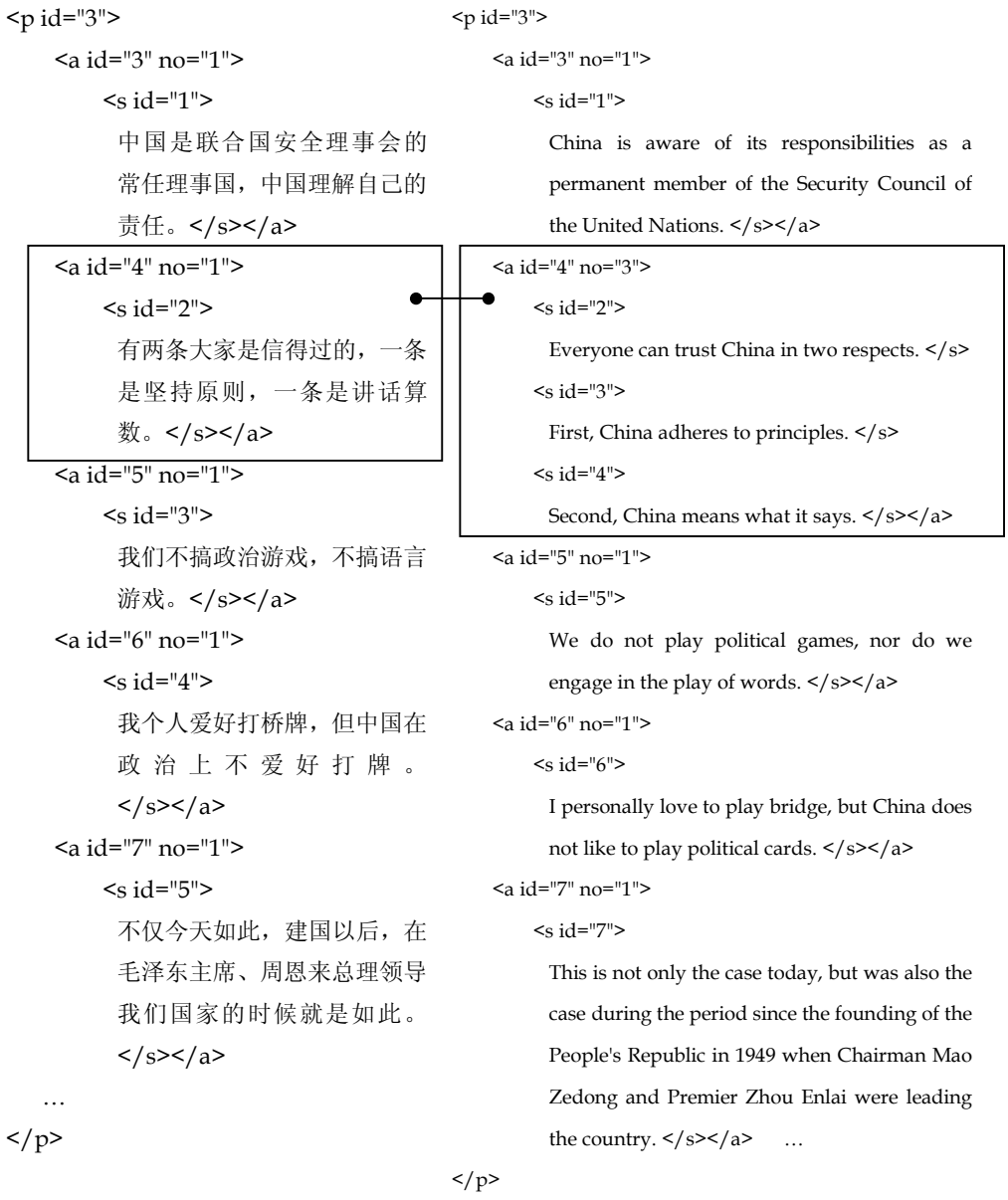

Figure-1 Fragment of the Chinese-English parallel corpus

2.4 The parallel corpus tool set

Corpus construction is very labour intensive; it cannot be done without automatic tools. To facilitate the construction of the Chinese-English parallel corpus, we have developed a set of corpus tools. So far we have the following tools in use: (1) the Chinese-English paragraph and sentence alignment program; (2) XML-based corpus encoding tool; (3) the Chinese segmentation and POS tagging program. The three tools have been heavily used in the construction of the Chinese-English parallel corpus as a pre-processing step before human verification. Figure-2 is a segmented and POS tagged fragment of the Chinese text.

```
<TEXT>
<TEXT_HEAD>
    <MODE>书面语</MODE><FIELD>工商</FIELD><STYLE>新闻</STYLE>
    <PERIOD>当代</PERIOD><CH_TITLE>今年中国经济和社会发展八项任务</CH_TITLE>
</TEXT_HEAD>
<TEXT_BODY>
    <p id="1">
   <a id="1" no="1">
    <s id="1">
      <CH_TITLE><w pos="t">今年</w>
            <w pos="ns">中国</w>
            <w pos="n">经济</w>
            <w pos="c">和</w>
            <w pos="n">社会</w>
            <w pos="v">发展</w>
            <w pos="m">八</w>
            <w pos="q">项</w>
            <w pos="n">任务</w></CH_TITLE></s></a></p>
     <p id="2">
```

Figure-2. A segmented and POS tagged fragment of the Chinese text

Apart from the above-mentioned tools, we also developed a simple English tokenization and lemmatization tool. A word level Chinese-English alignment is also under developing. But so far these tools have not been widely used in the corpus processing.

2.5 The workflow of the parallel corpus construction

To facilitate the construction of the parallel corpus, we also developed a systematic workflow based on our examination of the whole process of the corpus construction. According to the workflow, any text must be firstly processed in the following steps after they are collected and before they entered into the corpus.

1) Noise removing. Most of our texts come from Internet and therefore most of them are HTML files, there are many tags and irrelevant links in the text. These are irrelevant to the text contents and shall be removed from the texts in this step. After the first step, all texts are pure texts.

2) Textual attribute tagging. In this step, global textual attributes are tagged in the text. For example, the domain of the text is tagged in this step. This step is done with a special simple bilingual editor, which assists human to insert tags in the beginning of the text.

3) Parallel Alignment. Parallel alignment at paragraph and sentence level is done in this step with the alignment tools. For most of the texts, the alignment accuracy can be more than 98%.

4) Human verification of the alignment result. All the alignment results are verified and errors are corrected by human in this step. For the high accuracy of the alignment tools, human work is dramatically reduced.

5) Bilingual texts mark-up. With this step, the verified texts are encoded into XML format according to our mark-up guidelines. The XML mark-up is done automatically with XML encoding tools.

6) Segmentation and POS tagging of the Chinese texts. The Chinese part of the texts are automatically segmented and POS tagged with automatic tools. Due to the huge work required, the segmentation and POS tagged result is not verified by human, which might be done in later time. But we also believe that unverified segmentation and POS tagging are also helpful for many purposes.

7) Indexing the texts. The final processed texts are indexed and compiled for late use.

3 Concordancing the corpus: building tool for using the corpus

To facilitate research using parallel corpus in the field of bilingual dictionary compilation, language teaching and contrastive language study between Chinese and English, a parallel concordancer has been developed. It has a GUI interface and is compatible with the XML markup scheme.

3.1 Corpus indexing

Corpus search can be time consuming, especially when searching of frequently used words. To make the search faster, the corpus texts need to be indexed. Inverted index is used in the concordance tool. For the corpus is always growing in size and language researcher or lexicographers might be only interested with part of the corpus, so the index is designed to be incremental. New texts can be easily put into without reindexing the entire corpus as long as the new texts are properly encoded in XML.

3.2 Search for words or patterns in both languages

Searching can be either monolingual or bilingual, exact match or fuzzy match and either single word or a word pattern. User's searching intent could be specified using a mini-query language. Figure-3 shows the result of searching for Chinese word "民主" and English word "democracy", which lists all sentence pairs containing "民主" in the Chinese part and "democracy" in the English part, language researchers or translators could examine how many times the Chinese word "民主" is translated into English word "democracy".

. . . 党派阶级的**民主**政治的斗争. . .	...the struggle for **democracy** is bound to mani...
. . . 基本的是从**民主**政治斗争中. . .	...the struggle for **democracy** . That is...
. . . 必须贯彻**民主**的精神. . .	...a spirit of **democracy** in exercising lead...
. . . 封建的缺乏**民主**的国家，. . .	...country that lacks **democracy** , it is...
. . . 它能在**民主**政治斗争中. . .	...the struggle for **democracy** , the Party...
. . . 具有充分的**民主**精神，. . .	...Fourth , since **democracy** constitutes the ...
. . . 中的武断不**民主**的错误，. . .	...the struggle for **democracy** . First ,...
. . . 政治的开展，**民主**教育比任何. . .	..., education in **democracy** has become more...
. . . 我们要在**民主**政治斗争中，. . .	...the struggle for **democracy** we should ensure...

Figure-3. Result of searching for Chinese word "民主" and English word "democracy"

Language researchers or translators might not care the translation of a word, say "民主", or they want to investigate all possible translations, they can just query the corpus only with that word, then all sentence pairs with that word will be listed. In this way, they will find other possible translations of that word by checking the sentence translations, for example, " 民主" could be translated as "democracy" and "democratic" as well. It's also possible for a bilingual lexicographer to find a translation word that is not recorded in the existing bilingual dictionaries, so he might decide to include it in his dictionary project.

For all the Chinese texts are POS tagged, so the concordance tool enable searches for words with particular word class, for example, they could query the corpus for the occurrences of "翻译"[1] as noun or as verb. They can just type the word with the POS tag interested in the query box as "翻译/n".

Fuzzy search enables searching against all inflectional forms of English words; for example, one might be interested with all inflectional forms of word "take". He could search the corpus with fuzzy search.

[1] "翻译" , as a noun, means "translator" or "interpretor", means "translate" as a verb,.

Chinese is not a morphological language. Most of Chinese words do not have morphological variants. However, but different Chinese linguists might have different opinions when talking about whether a certain character string constitute a word. Thus in a segmented corpus, both "民主" and "民主集中制[2]" are segmented as word. Fuzzy search for "民主" could also list occurrences of "民主集中制" in the corpus.

Apart from the above-mentioned simple query on corpus, the mini-query language could be used to form complex query intention. Users can search patterns against corpus. For example, an English teacher might be interested in listing occurrences of pattern "take…into…account", which he could use them as class examples. Such search requirement could be accomplished with an expression like "*take+10into+10account" in the concordance tool, the number after "+" stands for the maximum distance between words in the expression. As an example for Chinese, the following query expression "*一+8就-" will list sentence pairs containing pattern "一…就…", with at most 8 words between "一"and "就", but without comma in between.

3.3 Sorting the search result

Language researchers maybe want to reorder the search result. They maybe want to group the resulted sentence pairs with similar context words. The parallel concordancer enables users to reorder the searching results against the context words of the search words. Figures-4 shows a sorting result against the pervious word of the search word "民主".

...没有无产阶级的**民主**和无产阶级的...
...工作和一系列的**民主**改革的任务...
...发展这样的**民主**政治斗争，因为它对于我们是有利无害的...
...这是一个重大的**民主**运动...
...发扬**民主**不会妨碍统一领导...
...集中制，发扬**民主**，加强集中...
...，在充分发扬**民主**的基础上，...
...怎么样，这对于发扬**民主**有好处...
...，这对于发扬**民主**，贯彻执行...
...真正地发展**民主**，民主政治斗...
...有意识地去发展**民主**政治的斗争，...

Figure-4. Sorting the search result

In Figure-4, sentences are grouped according to the first word left to the search word. Lexicographers can easily find the words with which the search word may collocate, such as "发扬"[3] and "发展", and this knowledge could then be depicted in his dictionary project.

3.4 Producing context profile for a search word

To facilitate research on collocations, a context profile could be produced for a search word, which is composed of the words collocates to the search item and frequencies of their co-occurrences. Figure-5 shows the context profile for search word "民主". In figure-5, **wd** column list words that "民主" co-occurred in sentences, **freq** column shows co-occurrences frequency. **Lx** (**Rx**) stand for the word is

[2] "民主集中制" means "democratic centralism".

[3] "发扬" means "promote" here, and "发展" means "develop".

the **x**-th word to the left (right) of the search word. Users can know "发扬" and "民主" co-occurred 5 times in the small sub-corpus loaded by the concordance tool.

L3		L2		L1		R1		R2		R3	
freq	**wd**	**freq**	**wd**	**freq**	**wd**	**freq**	**wd**	**freq**	**wd**	**freq**	**wd**
11	,	11	,	28	的	政治	28	的	21	,	20
6	民主	8	的	13	大	的	17	,	16	的	10
6	和	8	要	8	党内	政权	13	斗争	13	民主	6
5	也	4	是	6	在	,	12	不	4	中	4
5	有	4	充分	5	发扬	生活	12	基础	3	有	3
4	的	4	有	4	缺乏	和	6	上	3	敌占区	3
3	我们	4	、	4	有	政府	5	就	3	不	3
3	成立	4	党	3	不	、	5	精神	2	斗争	2
3	是	4	国家	3	民族	问题	3	作风	2	党	2

Figure-5. Context profile for search word

3.5 Producing frequency lists for both Chinese and English

Frequency list might be very useful for lexicographers. It could be used as measurement of the commonness of a word. For now, the parallel concordance tool could produce frequency list for both Chinese and English based on the texts loaded in the memory. For Chinese, frequency list can be produced with part of speech type. Figure-6 shows all three types of frequency list produced by the tool.

the	8830	的	7952	的	u-助词	7895
and	4572	是	2273	是	v-动词	2269
of	3990	不	1445	不	d-副词	1442
to	3619	在	1409	在	p-介词	1408
in	2863	我们	1358	我们	r-代词	1358
we	1816	了	1262	党	n-名词	1008
a	1727	党	1010	和	c-连词	990
party	1433	和	1005	有	v-动词	972
is	1252	有	974	要	v-动词	957
should	1119	要	958	了	u-助词	903

(1) English frequency list (2) Chinese frequency list (3) Chinese frequency list with POS

Figure-6. Frequency lists

3.6 Guessing translation word for a search word

Given the searching results are bilingual, a guessing mechanism based on hypothesis and test is implemented in the tool, which can list possible candidate translation words for a Chinese word or English word. Certainly, such guessing is not always 100% accurate, but it's helpful for the user to know the most frequently used translation of the search word. A translator might be interested with such information.

In addition, the parallel concordancing tool also provides operations to select, copy, paste or saving the search results, so that they could be used by the user in other situation.

4 Different use of the corpus

So far, the corpus has been being used for different purpose. Among them are:

(1) As the translation memory of a Monograph oriented Machine Aided Translation System. Actually, the original goal of building such corpus was providing translation examples for MT systems. Sentence and their translation could be easily extracted from the corpus and put into the translation memory. So far, 400,000 Chinese-English sentence pairs have been used in the above mentioned Monograph oriented MAT System.

(2) Contrastive study on Chinese and English passive form. BAI Xiaojing and ZHAN Weidong (2003) explored how English passive form could be translated into Chinese, either with Chinese preposition "被" or not. The conclusion is draw from observing the corpus.

(3) Using the corpus in the bilingual dictionary compilation. With the help of the concordancing tool, a joint effort to compile a medium-sized Chinese-English learner dictionary has been started. Also a workbench for lexicographers is also under development to help them in their work to build dictionary based on parallel corpus.

5 Conclusion

Chinese-English parallel corpus is certainly important resources for cross-language information processing, and also important resources for Chinese-English bilingual lexicography, language research and teaching. But how the parallel corpus could be used or effectively used in these areas is still worth being further investigated. This paper presents our works toward building Chinese-English parallel corpus and using the corpus. And the work is still unfinished. We hope we could continue to push forward both the corpus construction and tools development, so that the corpus could be effectively used for various purposes with the tools.

Acknowledgements

My thanks go to prof.Yu Shiwen, Dr. Bai Xaiojing, Dr.Zhan Weidong, Mr. Wu Yonghua, Miss Xiao Huayun, Mr. Zhang Huarui of the Institute of Computational Linguistics, Peking University, they all made their contributions to the work presented in the paper.

The work also gets support from Chinese Ministry of Education under a language research project, the project No. is YB-105.

References

BAI Xiaojing, CHANG Baobao and ZHAN Weidong (2002), The construction of a large-scale of Chinese-English parallel Corpus. In proceedings of National Machine Translation Conference 2002: the progress in Machine Translation, the Electronic Industrial Publisher, Beijing. pp.124-131.

CHANG Baobao, BAI Xiaojing (2003), The Markup Guidelines for the Chinese-English Parallel Corpus of Peking University, Journal of Chinese Language and Computing, vol.13,No.2, 2003, pp. 195-214.

CHANG Baobao(2003), Translation Equivalent Pairs Extraction Based on Statistical Measures, Journal of Computers, No.5, 2003, pp. 616-621.

BAI Xiaojing, ZHAN Weidong (2003), The constraint condition for Chinese BEI-sentence and its translation in Machine Translation, present in International symposium on Chinese passive expression, Wuhan, 2003.

Kwong, O.Y., Tsou, B.K., Lai, T.B.Y., Luk, R.W.P., Cheung, L.Y.L. and Chik, F.C.Y. (2001), A Bilingual Corpus in the Legal Domain and its Applications. In Proceedings of the NLPRS Workshop on Language Resources in Asia, Tokyo, Japan.