Bilingual Corpus Construction and its Management for Chinese-English Machine Translation[‡]

Chang-Baobao Zhang-Huarui Kang-Shiyong[†] Yu-Shiwen
The Institute of Computational Linguistics, Peking University, 100871
†The Department of Chinese Language and Literature, Yan Tai Normal College, 264025

Abstract: In recent years, mono-or multilingual corpora are viewed as key resources in language information processing and language engineering projects. To support an ongoing Chinese-English machine translation project, a Chinese English bilingual corpus is being set up. This paper gives a detailed discussion on construction of the corpus.

Keywords: Bilingual Corpus, Machine Translation, Corpus Markup, Corpus Annotation

1. Introduction

In recent years, mono-or multilingual corpora are viewed as key resources in language information processing and language engineering projects. Many new approaches to language-related applications and researches based on large-scale corpora are proposed. For Example, aligned bilingual sentence pairs could be directly used as a translation memory to improve the quality of the machine translation; Useful data or knowledge could also be extracted from bilingual corpus based on statistical model, such as acquisition of bilingual translation patterns. On the other hand, bilingual corpus could also be very valuable to bilingual lexicographers and linguists.

There are three inter-related fields concerning the study of multi- or bilingual corpus. (1) Techniques and methods to process or annotate collection of bilingual texts. Many proposals to tag, parse and align bilingual corpus were published and more and more programs or tools for such purpose have appeared [Gale 1993]; (2) Models based on multi- or bilingual corpus to specific applications in which multiple languages are involved. For example [Brown 1990] and [Nagao 1984] used bilingual corpus in different ways to develop machine translation system and [Klavans 1990] is interested in utilizing bilingual corpora in lexicography; (3) General issues in designing, compiling and encoding of corpora. TEI (Text Encoding Initiative) and CES (Corpus Encoding Standard) based on SGML are being developed to markup text structures. Most of researchers in China are focused their attention on the first two fields [Liu 1995], systematic and well-encoded bilingual corpora, especially with Chinese as source language, have been not available yet.

The Institute of Computational Linguistics in Peking University, National Key Laboratory for Intelligence Technology in Tsinghua University and the Institute of Computational Technology of Chinese Academy of Science are making a joint effort to develop a practical Chinese-English machine translation system funded by the Chinese government since January. 2000. For the purpose of combining all benefits of different translation methods, the system is designed as a

[‡] Supported by Chinese 973 High Technoloy Project (NO.G1998030507-4)

multi-engine system. In this paradigm, traditional rule based engine, corpus based engine and other translation engine will coexist and interact in the final system. As a key resource, a Chinese-English bilingual corpus with about one million Chinese characters and 0.6-million-word corresponding English texts is setting up. In this paper, the design, collection, markup and annotation of such a corpus are described in detail.

2. The design of the corpus

When compiling new bilingual corpus, even if it is small, it is sensible for compilers to design it carefully. Compilers must make decisions on the corpus type, size and composition. A valuable corpus is not arbitrary collection of arbitrary texts. Careful planning will ensure that large amount of work involved in compilation is worthwhile.

As to our task of constructing a bilingual corpus, we think it is very important to always keep intended usage of the corpus in mind. The corpus will serve to a Chinese-English machine translation system, which will mainly deal with newspaper news texts in Internet for assimilation purpose. We make it clear that our corpus is a specialized corpus instead of a general one. Highly coverage to all text types will make no senses, contrast to a corpus serving to linguistic analysis and research. The content of the corpus, the text categories, the structure of the corpus, the sources from which texts are collected, and the time when the texts are produced should fit for translating newspaper news. It would be very ideal if the corpus could be a sample of population of news texts under statistical meaning. However, it is very difficult to construct a corpus which could meet all theoretical requirements above mentioned. The first problem we face is that we do not have enough bilingual newspaper texts to collect and translation of Chinese newspaper is very expensive. We must make tradeoff between theoretical criterion and the availability of news texts. At last, we collect the text according to the following principle:

1) The text type should be news reportage at best, but some other materials similar to news texts with good translation will also be included in the corpus. Besides news reportage, we also collect some texts and its translation of press conference, some policy papers and its translation produced by Chinese government, and some essays with good translation.
2) All the bilingual texts should be with Chinese as source language at best, for serving to a Chinese to English machine translation system, but some existing easy-to-collect English sentences with professional Chinese translation also are introduced into the corpus. For example, about 25,000 English-Chinese sentence pairs derived from a machine translation evaluation project [Yu 1991], where such sentence pairs serve as test set, are included in the corpus.
3) All texts should be collected in full text, but there are also exceptions. We think full text will be a useful resource to learn text structure knowledge in the future.
4) All texts to be collected in the corpus should be published or produced in recent years.

Guided by these principles, we have collected about one million Chinese characters full texts and its English translations, mainly from Internet, and about 35,000 sentence pairs. All full texts are one of four types, --- news reportage, text of press conference, policy paper and essays. The composition of full text corpus is illustrated in Figure 1.

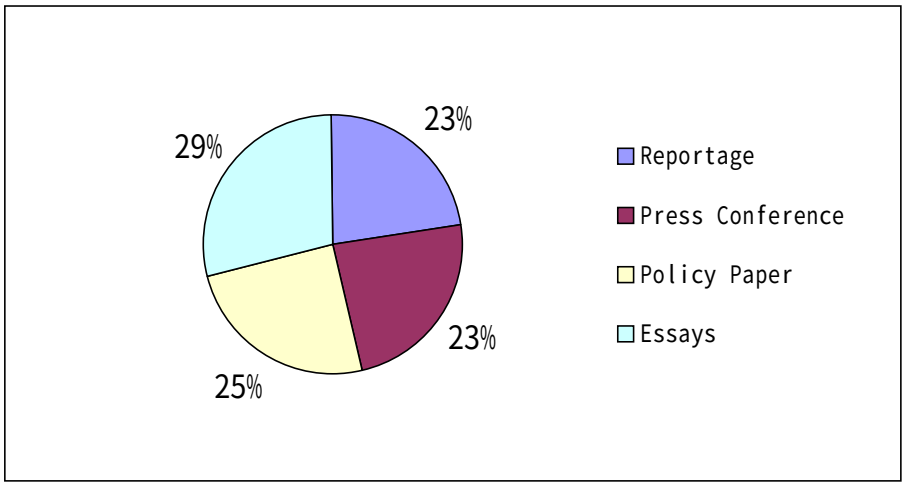

Figure 1. The composition of the full text corpus

3. The markup of the corpus

The best way to manage the parallel texts is to develop special tool programs, which will make the corpus easy to use. But that means that all texts from different sources must be encoded or markuped in the same way. A uniform representation could also make the corpus easy to interchange cross application and software platform. There are two famous corpus encoding standard proposals available, ---*Corpus Encoding Standard*, still under way to final version, and *Text Encoding Initiative*, which have been used by some monolingual corpus, such as *British National Corpus*. Both markup standard are based on *the Standard Generalized Markup Language* (SGML). Because most texts in our corpus come from the Internet website, most of them are originally encoded in *Hyper Text Markup Language*. We could also choose to develop new encoding system. Which one of four choices is the best way to markup could only be determined by careful consideration and comparison.

Firstly, HTML form was abandoned, even it seems least work is needed to encode all the texts. And it seems that downloaded text could be used directly without any modification. But that is not true. HTML is a widely used markup language by today's website and has many variations, different software enterprises, such as Microsoft, Netscape, made different extensions to it and put different new elements into it. The syntax of HTML is not strict, many web pages contains error in it and express the same meaning in different forms. Most importantly, HTML is presentation and content mixed markup system. There are both content tags, such as <Hn>, and presentation tags, such as <font>. Most of times, authors of pages do not use content elements for special display effect. For example, page authors prefer to use and <font> to make the titles of texts more eye-catching instead of using tags, such as <Hn>.

Both CES and TEI are designed to encode corpus. But the problem is that both of them are designed for general purpose. Both of them are difficult to grasp and use, even we only select a minimum set of necessary elements. Some of necessary tags according to them have less relevance to our purpose. But some elements to tag the structure of news reportage are not available. Furthermore, both of them are based on SGML, which is proven to be too complex to use and is not widely used by information technology society. To develop a fully functional SGML parser is not an easy task.

In order to achieve a simple, but workable with regard to our purpose, encoding solution, we choose to develop a new tagging system with reference mainly to CES. The new system does not try to cover all document types, but has necessary tags for document of news reportage, and only with general support to other document types. We also base our system on the most popular markup language nowdays, --- eXtensible Markup Language, which is a small subset of SGML and supported by many important software companys.

According to our encoding system, the whole corpus is composed of sets of inter-related documents. The logic structure of the bilingual corpus is shown in Figure 2.

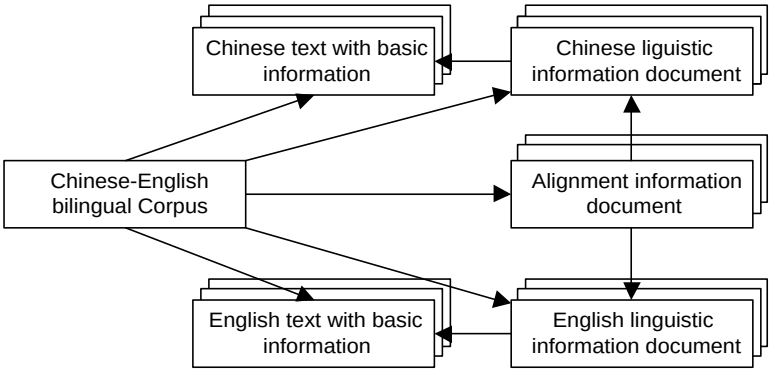

Figure 2. The logic structure of the bilingual corpus

In this paradigm, a Chinese text and its English Translation are represented by five cross-linked documents:

(1) Document containing the Chinese text with basic information;
In this Document, the structure of the original Chinese text is tagged with pre-defined tags. The title, subtitle, author, reportage introduction, communication head (such as "新华社，9 月 10 日北京电") and other elements of news reportage or general document are tagged. And named entity, such as person name, organization name, is also tagged in this document.

(2) Document containing the English text with basic information;
This document is similar to the first document, --- Document containing the Chinese text with basic information. The only difference is that this document is for English text.

(3) Chinese linguistic information document;
In this document, linguistic information about words, phrase, subsentence, sentence of the Chinese text are recorded.

(4) English linguistic information document;
In this document, linguistic information about words, phrase, subsentence, sentence of the English text are recorded.

(5) Alignment information document.
Alignment information about the Chinese text and English text is recorded in this document.

Four Document Type Definitions (DTDs) are defined for representing all document above mentioned. One DTD file is for general description of the whole bilingual corpus, one DTD file

for alignment information document, Document containing the Chinese text with basic information and Document containing the English text with basic information share one DTD file, and the last DTD file is for Chinese linguistic information document and English linguistic information document.

With the markup system, simple annotation and deep annotation can be tagged in a uniform way. And it also makes step-by-step annotation possible.

4. The annotation and alignment of the corpus

Concerning corpus annotation, the first issue is which annotations should be carried out on a corpus. That undoubtedly depends on how the final corpus will be used. Our corpus is supposed to be used in Chinese-English machine translation. We hope part of the corpus could be used directly or indirectly as resource in the Chinese-English machine translation system. We also hope useful translation knowledge and statistical data could be extracted from the corpus, such as bilingual lexicon, translation pattern. The second issue is there must be efficient way to annotate the corpus. Corpus annotation is very time-consuming, labour-intensive, expensive and sometimes error-prone.

At present, we are carrying out or planning to carry out the following kinds of annotations to the corpus.

1) Chinese word segmentation and pos tagging.
2) English tokenization and pos tagging.
3) Chinese and English proper noun identification are also planed, small experiment on Chinese organizational name was carried out.
4) Alignment of the Chinese text and English text at sentence level.
5) Alignment of Chinese proper noun and English translation are also planned.
6) Tagging the Chinese word with detail grammatical features with regard to the context. Such an annotation is based on the *Grammatical Knowledge-base of Contemporary Chinese* [Yu 1996]. In the knowledge base, words of some type have dozens of grammatical features and possible values. But with regard to context, only one or small subset of the values of features make senses in real text. Such annotation will make it possible to learn word translation rules. So far, we have conducted some experiment on such annotation.

Typical procedure to conduct the annotations is as follows: 1) If there are available program tools, annotate the corpus automatically firstly. 2) Verify the machine-tagged corpus by human professionals. With a Chinese word segmentation and tagging software, the Chinese text of the corpus is segmented and pos-tagged according to specification of the Institute of Computational Linguistics, Peking University. With an English tokenization program we developed and a modified freely downloadable pos-tagger, the English text are tokenized and tagged according to *Penn Treebank* tagset. An improved existing alignment program is used for alignment of the bilingual texts at sentence level. We also developed a rudimentary Chinese organizational name identification program based on statistical knowledge, but the experiment result does not show good performance (with recall= 69.7, accuracy = 40.9). Improvements are required before it comes into use.

So far, ten percent of the full text corpus are tokenized, tagged and verified. Verification on automatic sentence alignment result is under way. Figure 3 is samples of the annotated corpus:

[香港/ns 特别/a 行政区/n]ns 成立/v 以来/f ，/w 香港/ns 继续/v 保持/v 着/u 亚太地区/j 重要/a 的 /u 国际/n 金融/n 、/w 贸易/vn 、/w 信息/n 、/w 航运/n 中心/n 和/c 世界/n 最/d 大/a 的/u 自由港/n 地位/n ；/w 世界/n 最/d 大/a 的/u 集装箱/n 码头/n ——/w 香港/ns 葵涌/ns 货柜/n 码头/n ，/w 平均/a 每/r 天/q 有/v 112/m 艘/q 远洋/b 巨轮/n 进出/v 。/w 亚洲/ns 最/d 大/a 的/u 香港/ns 新机场/n ，/w 已/d 投入/v 运营/v ；/w 截至/v 今年/t 第一/m 季度/n ，/w 香港/ns 外汇/n 储备/vn 968 亿/m 美元/q ，/w 稳/ad 居/v 世界/n 第三/m 位/q ；/w 香港/ns 社会/n 人心/n 稳定/a 。/w 前/f 些/q 年/q 移居/v 海外/s 的/u 港人/n 大量/m 回流/v 。/w 今年/t 5 月/t ，/w [香港/ns 特区/n]ns 第一/m 届/q 立法会/j 选举/v ，/w 148 万/m 居民/n 踊跃/ad 参与/v ，/w 投票率/n 高/a 达/v 53%/m ，/w 超过/v 港/j 英/j 时期/n 立法局/n 的/u 投票率/n ；/w 香港/ns 社会/n 秩序/n 良好/a 。/w 据/p 香港/ns 警方/n 统计/v ，/w 特区/n 成立/v 以来/f ，/w 香港/ns 社会/n 的/u 整体/n 犯罪率/n 为/v 24/m 年/q 来/f 最低/a 的/u 一/m 年/q ，/w 全年/n 刑事/b 案件/n 比/p 上年/t 下降/v 了/u 15%/m 。/w 去年底/t 一/m 项/q 社会/n 调查/vn 显示/v ，/w 有/v 79%/m 的/u 人/n 感觉/v 在/p 香港/ns 生活/v 非常/d 安全/a 。/w

(1) Fragment of Tagged Chinese text

Since/IN the/DT founding/NN of/IN the/DT SAR/NNP ,/, Hong/NNP Kong/NNP has/VBZ maintained/VBN its/PRP$ status/NN as/IN an/DT important/JJ financial/JJ ,/, trade/NN ,/, information/NN and/CC shipping/NN center/NN in/IN the/DT Asia-Pacific/NNP Region/NNP ,/, and/CC the/DT largest/RBS free/JJ trade/NN port/NN in/IN the/DT world/NN ./. Kwai/NNP Chung/NNP Pier/NNP ,/, the/DT world/NN 's/POS largest/JJS container/NN wharf/NN ,/, berths/VBZ on/IN the/DT average/JJ 112/CD ocean-going/JJ vessels/NNS every/DT day/NN ./. Hong/NNP Kong/NNP 's/POS new/JJ airport/NN ,/, the/DT largest/JJS of/IN its/PRP$ type/NN in/IN Asia/NNP ,/, has/VBZ started/VBN operation/NN ./. By/IN the/DT end/NN of/IN March/NN, /, Hong/NNP Kong/NNP 's/POS foreign/JJ exchange/NN reserves/NNS had/VBD totalled/VBN US/NNP $/SYM 96.8/CD billion/CD ,/, steadily/RB ranking/VBG third/JJ Cin/IN the/DT world/NN ./.

(2) Fragment of Tagged English text

{{{{
[[(1:1)
(%s)第一，在人民币不贬值的条件下，立足自己，发挥优势，降低成本，增加出口的竞争能力。
%$

(%s)First, under the condition that the Renminbi (RMB) will not devaluate, we will rely on own efforts and advantages to lower cost and increase the competitiveness of our exports.

]]

[[(1:1)

(%s)降低成本主要是依靠自己，特别是依靠出口企业从过去的粗放经营向集约化经营转变，这方面潜力很大。

%$

(%s)In this regard, we will chiefly rely on transforming the management of export-oriented enterprises from an extensive to an intensive pattern.

]]

}}}}

(3) Fragment of aligned texts

Firgure 3. Samples of annotated corpus

5. Further work and application of the corpus

Surely, there are many further works to be done to the corpus, such as further deep annotation, which is not planned given a situation of lacking supporting tools and its complexity.

After 10 percent of the corpus was tokenized, tagged and aligned at sentence level, the task of tokenizing, tagging and aligning other 90 percent of the corpus is underway.

There are many ways to use the corpus in Chinese-English machine translation system. We only show how a translation memory engine we are developing uses the corpus.

Bilingual sentence pairs are key resource of translation memory engine. The sentence pair could be either simply annotated or deeply annotated, at present, the TM engine supposed to have two different levels of sentence annotation. In the first level, Chinese sentence and English sentence are aligned, but the sentences are not tokenized and tagged. For example, alignment of the ten percent of the full text corpus produce 2,534 sentence pairs, these sentence pairs are being indexed to form a simple translation memory in which there is no other information besides alignment information. With such translation memory, TM engine only provide search operation. If the input of TM engine happens to be in the memory, the output translation will be produced directly.

In the second level, Chinese sentence and English sentence are aligned, but the sentences are also tokenized and tagged, furthermore, Proper noun are also aligned. Such a collection of sentence pairs also could be indexed to form a translation memory. Because the translation memory contains much more information than simple translation memory, TM engine will view the sentence pair as sentence level translation pattern with the aligned proper nouns as pattern slots. For example, Sentence pair (1) will correspond to pattern (2).

(1)应 克林顿 总统 的 邀请 ,中国 总理 朱镕基 将 于 4月 6日 至 14日 访问 美国 。

At the invitation of President Clinton , Chinese Premier Zhu Rongji will visit the United States between April 6 – 14 .

(2) 应 X1 X2 的 邀请 ， X3 X4 X5 将 于 X6 X7 至 X8 访问 X9 。

At the invitation of X2’ X1’ , X3’ X4’ X5’ will visit X9’ between X6’ X7’ – X8’ .

Such sentence level translation patterns give TM engine ability to substitute proper nouns. For

Example, the TM engine will perform the following tranlation by utilizing the above patern and bilingual lexicon:

INPUT: 应 江泽民 主席 的 邀请 ，马来西亚 总理 马哈蒂尔 将于 10 月 3 日 至 6 日 访问 中国 。

OUTPUT: At the invitation of President Jiang Zemin , Malaysian Premier Mahathir will visit China between October 3 – 6 .

6. Acknowledgments

I would like to thank Prof. Duan Huiming, Dr. Chen Yuzhong, Dr. Wu Yunfang, Prof. Liu Qun and some other colleagues and students in the Institute of Computational Linguistics, Peking University for their contribution in verifying the machine-tagged corpus. Thanks also should be given to Dr. Wang Bin of the Institute of Computational Technology of Chinese Academy of Science, who provides initial program for aligning the corpus at sentence level. Improving an existing tool than developing a new one saves much work. It is also should be noted that such a work could not be done by authors of the paper alone. Thanks should also be given to any one who was or is helpful in this work but whose name is not listed above.

References

[Brown 1990] Brown, P., et al, A statistical approach to machine translation, Computational linguistics, V16, No.2, 1990

[CES] Corpus Encoding Standard, http://www.cs.vassar.edu/CES/

[Gale 1993] Gale W., et al, A program for aligning sentence in bilingual corpora, Computational linguistics, V19, No.1, 1993

[Klavans 1990] Klavans, J., and Tzoukermann, E., The BICORD system, In Proceedings, 15th International Conference on Computational Linguistics.

[Liu 1995] Liu Xin, Zhou Ming, Huang Changning, Experiment on alignment algorithm for Chinese-English parallel texts based on sentence lengths, In Chen Liwei eds. Progress and application in computational linguistics, Tsinghua University publish house, 1995 (in Chinese)

[Nagao 1984] Nagao, M., A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, In: A.Elithorn et al eds. Artificial and Human Intelligence, NATO Publication

[TEI] TEI Guidelines for Electronic Text Encoding and Interchange, http://etext.virginia.edu

[Yu 1991] Yu Shiwen, Jiang Xin, Zhu Xuefeng, Automatic system for evaluating machine translation, In: Proceedings of MMT’91, 1991(in Chinese)

[Yu 1996] Yu Shiwen, Zhu Xuefeng, The specification of The Grammatical Knowledge-base of Contemporary Chinese, Journal of Chinese information processing, 1996, No. 2