Compilation of A Multilingual Parallel Corpus

Yasuhito Tanaka
Hyogo University

Address:2301 Shinzaike Hiraoka−Cho
Kakogawa Hyogo
675−0101 Japan
Tel:+81−794−27−9940 Fax:+81−794−27−5112
E−mail: yasuhito@humans−kc.hyogo−dai.ac.jp

[0] Introduction

The author has already conducted a fair amount of research into how to compile a multilingual parallel corpus and has experimented with a few compilation methods. Furthermore, his findings on the characteristics of these methods and those of corpora created by them have been published.

In this paper, the author proposes a new method based on the extraction of data from Japanese−English bilingual newspaper articles and broadcast media news reports published on the WWW.

[1] Matters to Be Considered Before Compiling a Multilingual Parallel Corpus

The following matters must be clarified before compiling a multilingual parallel corpus:

(1) Languages of interest

Reasons for this as well.

(2) Specialty area

Examples include politics, economics, information technology and life sciences.

This issue is intertwined with the intended purpose of the corpus.

(3) Compilation method

Although researchers tend to focus on this aspect because of their natural interest in technical methodology, the compilation method should be examined after items (1) and (2). Nevertheless, it is important to find an efficient compilation method given the limited financial and human resources available, while always keeping the issue of copyright in mind.

[2] Past Compilation Method

One method the author used in the past to compile a multilingual parallel corpus is to have a number of students enter about 300 data items each, followed by duplication removal and entry error correction operations. After three to four years of efforts, the author was able to produce a Japanese−English bilingual parallel corpus consisting of some 212,000 sentence pairs. Classifying and sorting the sentence pairs by the number of words contained in the English sentence (length of the English sentence) yielded some interesting information.

Overall, the corpus had the following characteristics:

1) 40% of the sentences had a personal pronoun as the subject.

2) The average length of English sentences was 7.72 words, with the longest being 45 words.

3) Most often, sentences were everyday−use sentences.

4) Interrogative and exclamatory sentences accounted for only 7.64% and 0.95%, respectively.

These characteristics differ significantly from those of sentences appearing in newspaper or magazine articles or other similar materials. Journalistic sentences rarely have a personal pronoun as the subject and use a different group of verbs. They are also generally long. The author will now focus on this class of sentence.

[3] Data from Newspaper Articles and Other Similar Sources

Data prepared through translation

One way of compiling a bilingual parallel corpus from newspaper articles is to have them translated by translators.

The natural language processing group of a major Japanese company is said to be using 10,000 translated sentences in various fields originally extracted from newspaper articles as experimental data. Assuming a unit translation rate of 1,000 yen per sentence, this exercise must have cost the company 10 million yen, a prohibitive figure for academics like the author.

There are other problems with this approach as well. The use of a single translator may result in an idiosyncratic bias in translations, while the participation of multiple translators would give rise to a problem with the standardization of terms.

One may think of machine translation, but it is still not sophisticated enough to translate long sentences accurately. For this reason, the author has decided to turn to a new method, the extraction of bilingual sentence data from newspaper articles and broadcast media news reports published on the WWW.

[4] Matters to Be Considered Before Utilizing WWW

Massive amounts of text data are available on the WWW, and there are numerous Web pages with multilingual content. Prof. Kenji Kita of the University of Tokushima and others have succeeded in retrieving the Internet addresses of multilingual−content Web pages based on a morphological study of Web pages. This is a significant development. The author hopes that language combination information (Japanese−English, Japanese−Chinese, Japanese−French, and so on) will also become available in the future. This paper will discuss the Japanese−English combination.

[5] Compilation of Bilingual Parallel Corpus from News Data Accessible on WWW

The author decided to use those bilingual newspaper articles and broadcast media news reports published on the WWW that have clear correspondence between Japanese and English sentences.

However, even such materials are not uniform. Here are some typical variations.

(1) For reports originating in Japan, Japanese sentences are very simple, while the corresponding English sentences are more detailed and elaborate.

(2) Correspondence between Japanese and English texts is sometimes not exact, with the difference particularly pronounced in the order of the presentation of facts.

(3) With news reports published by CNET, a major US media company, English is the dominant language. Nevertheless, correspondence between English and Japanese sentences is good because Japanese translations are generally true to English originals.

In some cases, sentence−level, one−to−one correspondence is not so strict, but this is a relatively minor problem.

CNET reports are also translated into other languages, and the standard of translators is high. For these reasons, the author gave his students the following URLs and had them extract 200 items of data from those sites:
http://japan.cnet.com (Japanese)
http://home.cnet.com (English)

Task Create a file containing Japanese and English sentence pairs
The data format is as follows:

1. Number (in half−size characters)
2. Japanese sentence (in full−size characters)
3. English sentence (in half−size characters)

Some of the extracted sentence pairs are shown below.

¥1¥0001
¥2￥ワシントン州レッドモンド発―ソフトウェアメーカー大手、米マイクロソフトは14日（米国時間)､同社の新しい商用オペレーティング・システム（OS)、「Windows 2000」の販売数が100万本を突破したと発表した。マイクロソフトは同製品を、「社運を賭けた」製品と呼んでいた。
¥3¥ REDMOND, Wash.−Microsoft said today it had sold more than 1 million copies of Windows 2000, the software giant’s new operating system for a business that it has called a “bet the−company” product.

¥1¥0002
¥2￥この数字には先月市場に登場した3つのバージョン―デスクトップ用の『Professional』に加えて、『Server』､ウェブサイトや企業ネットワークを運営するための強力な企業向けコンピューター用の『Advanced Server』―の世界中の販売数が含まれている。
¥3¥ The figure included worldwide sales for each of the three versions that hit the market last month: Professional for desktops, and Server and Advanced Server for the powerful corporate computers that run Web sites and business networks.

¥1¥0003
¥2￥マイクロソフトによると、この販売総数には小売りアウトレット、流通チャンネルの再販業者、パソコンメーカーによる販売数が含まれているが、企業ライセンスを通じてこのソフトウェアを利用している大手ユーザーへの販売数は含まれていないという。
¥3¥ Microsoft said the sales total included those sold through retail outlets, channel resellers and personal computer makers but not through big customers that were deploying the software through enterprise licenses.

When extracting data, be aware that sentences may contain line break characters (⤵). These characters must be removed, as they will cause problems in the following step of processing (editing).

The author gave students ample time to do the task. Nevertheless, there were data duplications because of simultaneous student access to the same sites. These duplications were subsequently removed.

For students, it was meant to be an exercise to get used to the Internet and collect news data from it. Because of copyright implications, it was decided to use the collected sentence pairs as the author’s personal experimental data.

[6] Data Collection

The data collection results are presented below.

(1) Number of students 96 (each students 200 sentences)
(2) Number of data items collected 19,200
(3) Number of duplications 3,000
(4) Net number of data items obtained 16,000

The experiment made it possible to evaluate the amount of work required to collect data and the actual procedure of the new data collection method.

[7] Sorting Data

The experiment demonstrated the feasibility of the new data collection method.

The author plans to sort the collected data in the future using the following methods:

(1) Ordering sentence pairs according to the length of the English sentence and studying frequencies of sentence occurrence by the length of the English sentence.

(2) Checking correspondence between English and Japanese sentences.

It is necessary to check the data for any errors in the correspondence of English and Japanese sentences that may have resulted from the students' in adequate English proficiency or erroneous copying.

(3) Additions or omissions to data

Even when there is overall correspondence between English and Japanese sentences, the data must still be checked for any additions or omissions made to either sentence, as such additions or omissions are sometimes employed by publishers to compensate for differences between the readers of Japanese and English sentences, in terms of familiarity with the topic, degree of interest, relative importance, etc.

For example, the simple expression "Microsoft" in English sometimes becomes " Microsoft Corp., a major U.S. software firm" in Japanese.

This gives rise to the need to address the difficult question of what the equivalence of sentences between different languages means.

[8] Conclusion

Using a new method, the author was able to obtain 16,000 data items. The experiment showed that the new experimental data can be collected quite cheaply. However, any copyright problems must be resolved through negotiations with the copy−rights holders, typically newspaper publishers and broadcasting companies, with an appropriate contract signed as necessary. The data obtained in this experiment will be used exclusively for the author's personal research purposes.

[9] Bibliography

1) Tanaka, Yasuhito, Evaluation and Improvement of Machine Translation Systems, SIGNL Conference Paper 133−1, Information Processing Society of Japan, p.p. 1−6, 1999

2) Boguraev, Branimir and Pustejovsky, James, Corpus Processing for Lexical Acquisition, MIT Press, 1996

3) Armstrong, Susan, Using Large Corpora, MIT Press, 1994

4) Kita, Kenji and Yamaguchi, Naohiro, Automatically Compiling Multilingual Translation from the World Wide Web, SIGNL Conference Paper 128−18, Information Processing Society of Japan, p.p.127−134