[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

JEIDA's English-Japanese Bilingual Corpus Project

Hitoshi ISAHARA
Communications Research Laboratory
588-2 Iwaoka, Iwaoka-cho, Nishi-ku, Kobe 651-2401 JAPAN
[isahara@crl.go.jp]

Abstract

JEIDA (Japan Electronics Industry Development Association) has been developing a large bilingual aligned corpus for research in NLP, since the 1996 Japanese fiscal year. In fiscal year 1996, JEIDA did a feasibility study and received permission from the Japanese Ministries to create such a resource. JEIDA, then, made a "small" sentence aligned corpus in fiscal year 1997. JEIDA's new project started this April (1998) is aimed at developing a much larger corpora with more precise tags.

Introduction

A huge amount of bilingual data is necessary for NLP (natural language processing) research, for example, corpus-based research on MT systems. It can be used to extract fundamental data for research and to verify the research results.

There are some such corpora for Indo-European languages. However, there is no such bilingual corpus for Japanese and other languages that is generally available for research purposes. This is due to the problem of copyright and the cost of data development. Therefore, JEIDA (Japan Electronics Industry Development Association) decided to develop its own bilingual (English-Japanese) corpus for NLP research and make it publicly available without charge.

JEIDA is a joint organization of computer-related companies in Japan. The committee on text processing technology is a subcommittee of JEIDA's natural language processing committee. This subcommittee developed JEIDA's testsets for the quality evaluation of machine translation systems (Isahara, 1995). It has been developing a large bilingual aligned corpus for research in NLP, since the 1996 Japanese fiscal year.

In fiscal year 1996, we did a feasibility study and received permission from the Japanese Ministries to create such a resource. We, then, made a "small" sentence aligned corpus in fiscal year 1997. A new project started this April (1998) is aimed at developing a much larger corpora with more precise tags. An overview of this bilingual corpus project is presented in this paper.

The Source of our Corpus

We first decided on the source documents. We selected white papers from Japanese Ministries. The reasons for this choice are:

(1) both Japanese versions and their English translations exist,
(2) governmental papers have fewer copyright problems than commercial publications, and
(3) white papers cover a wide range of topics.

Because of (1), the English sentences in these white papers would not be considered "good" translations of the meaning in context but are merely sentence-to-sentence or paragraph-to-paragraph translations. However, they, therefore, suit the current state of NLP research. We have already gotten permission to use white papers from three Japanese ministries: the Environment Agency, the Economic Planning Agency and the Science and Technology Agency. We have developed an aligned bilingual corpus using six white papers from the 1992 to 1996 fiscal years (Table 1) and are now enlarging it with six other white papers.

We are trying to get permission to use another kind of document, such as monthly journals and manuals. This would make our corpora a kind of "balanced" corpus.

Ministry	Japanese Version	English Version
Environment	Heisei 6th	1993-1994
Economic Planning	Heisei 7th Heisei 8th	1994-1995 1995-1996
Science and Technology	Heisei 6th Heisei 7th Heisei 8th	1994 1995 1996

Table 1: Source of aligned bilingual corpus

Electronization and SGML Tagging

Some of the white papers are available on CD-ROMs or floppy disks and others are available only in a printed form. The latter, we had to input either manually or by using an OCR. The size of each document is shown in Table 2.

We are formatting our corpus in TEI format using the following steps:

(1) Definition of document type.
We define the document type of our bilingual corpus based on the TEI Lite regulation and its extensions. For chemical formulas, we adopted STANDCOM.DTD in ISO/IEC TR 9573-11. (Burnard, 1995; Bonhomme et al, 1995; Maler and Andaloussi, 1996)

(2) Conversion of nonstandard characters.
Gaiji (nonstandard characters) in Japanese, are converted into some combinations of standard characters. For example, "1 in a circle" is converted into "&c-l;".

White Paper		Size (byte)	Section	Paragraph	Sentence
Environment (Heisei 6th)	Japanese	1,175k	693	2,100	4,525
Environment (1993-1994)	English	1,535k	693	2,238	6,432
Economic Planning (Heisei 7th)	Japanese	601k	332	1,291	3,080
Economic Planning (1994-1995)	English	741k	332	1,279	3,645
Economic Planning (Heisei 8th)	Japanese	520k	339	816	2,761
Economic Planning (1995-1996)	English	766k	339	824	3,265
Science and Technology (Heisei 6th)	Japanese	417k	289	948	1,738
Science and Technology (1994)	English	655k	289	1,307	2,471
Science and Technology (Heisei 7th)	Japanese	434k	326	967	1,881
Science and Technology (1995)	English	689k	326	1,277	2,695
Science and Technology (Heisei 8th)	Japanese	383k	254	828	1,630
Science and Technology (1996)	English	604k	254	944	2,375

Table 2: The size of each white paper document

(3) Regularization of titles and bodies.
Before tagging bilingual texts, we have to regularize the texts so that we can identify their titles and bodies automatically. We did this regularization process manually because the titles in the English versions tend to be very different from the titles in the Japanese versions.

(4) SGML tagging.
After the regularization, most parts of the tagging, e.g., (a) identification of the hierarchy of sentences, (b) identification of titles, (c) identification of paragraphs, and (d) identification of sentences, can be done automatically. We are using only part of the tags defined by TEI Lite, e.g., tei, teiHeader, text, body, div, head, p, s, and q. Tasks which we have to do manually, e.g., assigning alignment attributes and identification of quotations, still remain to be done.

As for the character code, this bilingual corpus utilizes JIS (Japanese industrial standard) X 201 and JIS X 0208 for the Japanese text and JIS X 201 for the English text. They can be easily converted into EUC code. For Entity Sets, we utilize public entity sets such as ISOlat1, ISOgrk3, ISOpub, ISOnum, and ISOamsr. We have, also, defined our own entity set, as follows:

<!ENTITY amp CDATA "&" -- ampersand --> &
<!ENTITY It CDATA "<" - less-than--------> <
<!ENTITY gt CDATA ">" -- greater-than --> >

Problems we faced during our corpus development include the following;

(1) Identifying quotations
(a) Quoted isolated sentences
(b) Quotations embedded in sentences
(c) Remarks embedded in sentences

(2) Identifying itemized sentences
(a) Ordinary itemized sentences
(b) Itemized sentences embedded in another sentence
(c) New lines within itemized sentences

(3) Representation of chemical formulas
(a) Chemical terms which can not be represented by ordinary fonts
(b) Chemical terms which can be represented by ordinary fonts only

These problems are still unsolved, therefore, we have to process them manually.

Sentence to Sentence Alignment

In the 1997 fiscal year, we are aligning Japanese sentences with English sentences. Alignment data is a set of one-sentence to one-sentence correspondence from two sets of sentences extracted from corpora in Japanese and English. When one Japanese sentence (J1) is translated into several English sentences (E1, E2, , Ek), the correspondence is represented as {(J1, E1), (J1, E2), ..., (J1.Ek)}.
Aligned data is developed via automatic processing by using an alignment tagger and by manual postediting. Automatic processing is done by software developed by NTT Communications Laboratory (Haruno, 1997). Tools for the postediting of bilingual data and for data conversion have been developed by the committee. The postediting tool has a graphical user interface to make the process of postediting efficient. In our experience, postediting takes one minute per sentence.

Alignment Process and Data Format about Links

The input to the alignment process is an SGML tagged parallel corpus as described in the above section. The conversion tool removes unnecessary tags from the corpus and converts tags for sentence and paragraph delimiters into a format suitable for the automatic alignment tagger. The tagger analyzes the input and generates alignment data automatically. Sentences are divided differently in the SGML tagged corpus and the automatically aligned corpus. Therefore, a conversion is done to adjust them. Next, using the postediting tool, automatically-aligned data is postedited manually, to correct correspondences between sentences in Japanese and English. This tool is activated by a data pair, i.e., Japanese text and English text. Figure 1 shows a display of the postediting tool. In the center of the display, Japanese text is shown on the left side and English text on the right. Correspondence relations are represented via lines between them. Postediting is done by adding lines, deleting lines and changing attributes of lines. There are three attributes, i.e., automatically aligned, manually added and manually

removed. These attributes are represented by different line colors. The posteditor can add comments on the information obtained during postediting. Data about links and comment data are saved into files when the postediting is complete. Finally, postedited aligned data is converted into an SGML format. Correspondence information is tagged "TRANSLATION". This has two possible attributes, i.e., "FROM" and "TO". For example, when (the whole or part of) the Japanese sentence "J2.1.1.4-1.1" is translated into (the whole or part of) the English sentence "E2.1.1.4-1.1", the relationship is represented as <TRANSLATION FROM="J2.1.1.4-1.1" TO="E2.1.1.4-1.1 ">. The aligned data is a set of these data.

A Description of the Postediting Tool

The postediting tool is written in Tcl/Tk and run on UNIX or Windows 95/NT with Japanese Tcl/Tk. The tool functions as follows;

(1) It adds, deletes and changes attributes of correspondence.
These functions are done by "dragging" with the mouse.
(2) It makes comments on the correspondence and/or the contents of the text. (Figure 2)
Posteditors can save as comments, information on ellipsis, free translations, and translation pairs discovered during the postediting process. Comments can be added to one correspondence link, several links and text as a whole.
(3) It moves the cursor to the comment position.
(4) It saves the entry with or without comments.
(5) It prints out the whole piece of text or part of text around the cursor or it saves the print image into a postscript file.

Future Directions

Our corpus has only SGML tagged text, alignment data and manually added comments. We will try to add more precise tags to our bilingual corpus, such as part-of-speech tags, word alignment tags, clause alignment tags and syntactic and semantic tags.
For word alignment tags, we plan to add correspondences between proper nouns in Japanese and English. This kind **of** information is written by the posteditor manually in a comment now, to use this (for example) as test data for an information retrieval system, this information must be stored in a fixed format. However, problems of exhaustiveness still remain. We have to define what "proper noun" is objectively and plural number of manual check must be necessary. Since the automatic alignment tagger generates correspondence data during the tagging process, we can use this information as a hint for the posteditor.
Using the correct correspondence of proper nouns postedited by hand, the automatic tagger would be able to re-tag the original parallel text more precisely, and this correspondence could be used to improve the quality of the automatic tagger.
The easier plan is to tune the automatic alignment tagger to white papers. White papers have distinguishing characteristics, such as many numerical formulas, many proper nouns and many abbreviations. Using the correspondences of these constituents, will improve tagging efficiency.
Considering clause alignment tags; clause level correspondences are useful for MT research, because more general linguistic information can be extracted from these than from sentence alignment tags. This can be used as fundamental data for example based machine translation systems.
Also, our plans include the enlargement of the bilingual aligned corpus. We aim to develop a corpus ten times bigger than the one that we have now.

Conclusion

In this paper, we have discussed JEIDA's bilingual corpus project. This corpus is being developed to be:

(1) available without charge to the public for research and evaluation of NLP technology,
(2) built under cooperation and dispersion, and
(3) general and independent of any one specific linguistic theory.

We will continue our efforts to enlarge this public bilingual aligned corpora for NLP research on these principles.

References

Bonhomme P. et al.(1995). LINGUA Information & Technical Aspect. Lingua Project.

Burnard, L. (1995). TEI Lite: An Introduction to Text Encoding for Interchange. C. M. Sperberg-McQueen.

Haruno, M. (1997). Bilingual Text Alignment Using Statistical and Dictionary Information (in Japanese). *Transactions of Information Processing Society of Japan,* 38(4).

Isahara, H. (1995). JEIDA's Test-Sets for Quality Evaluation of MT Systems - Technical Evaluation from the Developer's Point of View -, In *Proceedings of the MT Summit V.*

Maler E. and J. El Andaloussi (1996). *Developing SGML DTDs From Text to Model to Markup,* Prentice Hall PTR.

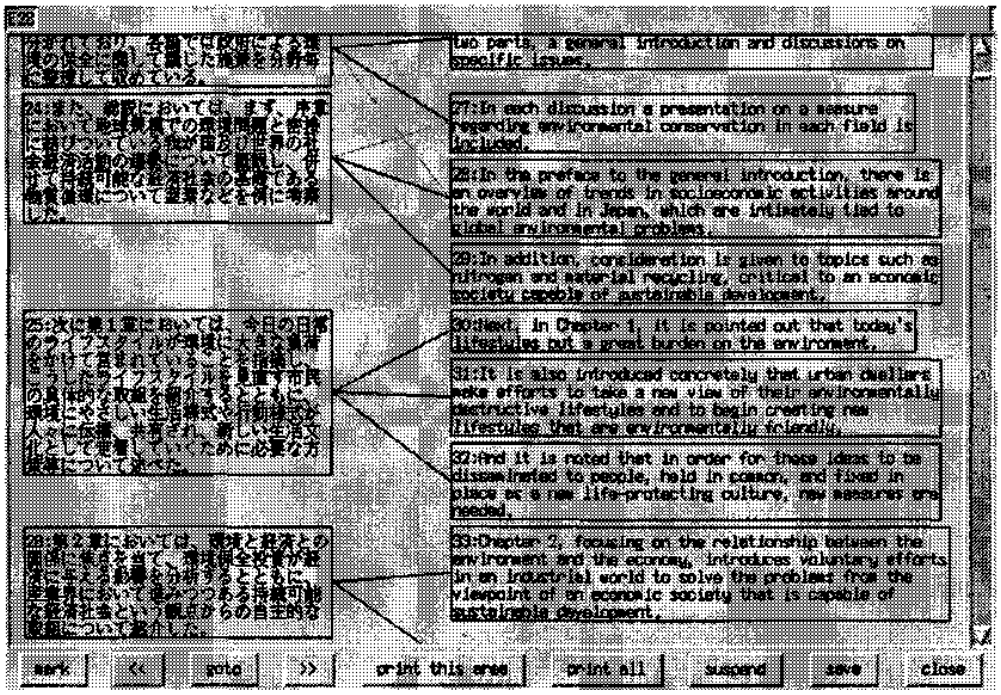

Figure 1: The postediting tool for a bilingual aligned corpus

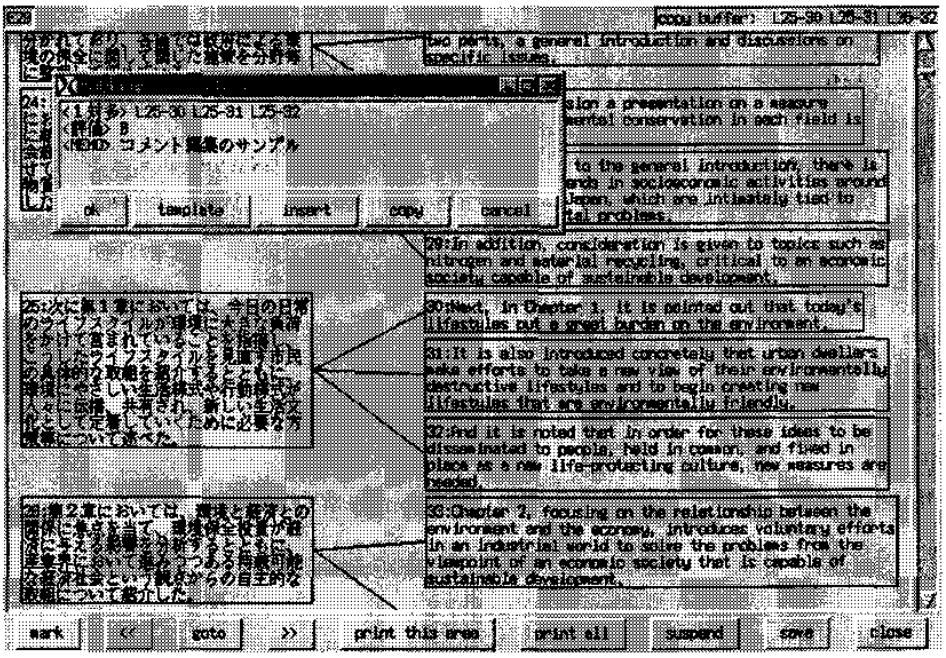

Figure 2: The comment edit window