JEIDA's Bilingual Corpus and Other Corpora for NLP Research in Japan

Hitoshi ISAHARA

Intelligent Processing Section
Communications Research Laboratory
Ministry of Posts and Telecommunications

588-2 Iwaoka, Iwaoka-cho, Nishi-ku, Kobe
651-24 JAPAN
Email: isahara@crl.go.jp

Abstract

The committee on text processing technology of JEIDA (Japan Electronics Industry Development Association) has been developing its bilingual corpus for research on machine translation systems since the 1996 Japanese fiscal year. An overview of this bilingual corpus is presented in this paper. And other linguistic data recently developed in Japan, which includes the RWC text database and the simple sentence data by the CRL and IPA.

1. Introduction

The committee on text processing technology of JEIDA (Japan Electronics Industry Development Association) is a subcommittee of JEIDA's committee on natural language processing (chairman: Prof. Nagao of Kyoto University), and developed JEIDA's testsets for the quality evaluation of machine translation systems [1]. The committee has been developing its bilingual corpus for research on machine translation systems since the 1996 Japanese fiscal year.

An overview of this bilingual corpus is presented in this paper. And other linguistic data recently developed in Japan, which includes the RWC text database and the simple sentence data by the CRL and IPA.

2. JEIDA's bilingual corpus

A huge amount of bilingual data is necessary for MT research, especially for corpus-based research on MT systems. There are some such corpora for Indo-European languages, however, there is no such bilingual corpus for Japanese and other languages which is available to the public for research purposes. Therefore, we decided to develop our own bilingual (English-Japanese) corpus for MT research.

We started with the decision on the source documents. We selected white papers from Japanese Ministries as the source of our corpus. The reasons why we chose white papers 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), 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 and therefore, they suit the current state of NLP research. Our plan to do research on this bilingual corpus includes research on alignment of parallel texts and acquisition of linguistic information from parallel texts.

We have already gotten permission to use white papers from three Japanese ministries: the Environment Agency, Economic Planning Agency and Science and Technology Agency. We are using white papers from the 1992 to 1996 fiscal years. Some of them are available on CD-ROMs or floppy disks and others are available only in a printed form, which we had to input manually or by using an OCR.

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. As for chemical formulas, we adopted STANDCOM.DTD in ISO/IEC TR 9573-11. (See [2], [3], and [4] for these regulations.)

(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;".

(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) Tagging.

After the regularization, most parts of the tagging, e.g., (a) identification of 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. However, tasks which we have to do manually, e.g., assigning alignment attributes and identification of quotation, still remain.

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

```
<!ENTITY amp CDATA "&#38;"  -- ampersand -->     &
<!ENTITY lt CDATA "&#60;"    -- less-than----->    <
<!ENTITY gt CDATA "&#62;"    -- greater-than -->    >
```

In this fiscal year, we are aligning Japanese sentences with English sentences using an alignment tool developed by NTT. And, we will also try to add more precise tags to our bilingual corpus, such as word alignment tags, part-of-speech tags, and syntactic and semantic tags. Figure 1 shows an example of our corpus (from a “White Paper on the Environment”).

Japanese	English
<div4 type=subsection id="J2.1.1.4">	<div4 type=subsection id="E2.1.1.4">
<head id="J2.1.1.4-h" corresp="E2.1.1.4-h"> (4) 農林水産物の生産と消費の増大 </head>	<head id="E2.1.1.4-h"> (4) Expansion in the production and consumption of agricultural, forest, and marine products </head>
<p id="J2.1.1.4-1" corresp="E2.1.1.4-1">	<p id="E2.1.1.4-1">
<s id="J2.1.1.4-1.1" corresp="E2.1.1.4-1.1 E2.1.1.4-1.2"> 農林水産業は、食糧や木材等の供給により人類の生存を最も基礎的なところで支えている重要な活動であり、また、農林水産業が営まれている地域においては、適切な農林水産活動を通じて農地、森林等が有する環境保全能力が維持されている。</s>	<s id="E2.1.1.4-1.1">By supplying food and timber products, the agriculture, forest, and marine products industries provide the most basic support for human existence.</s> <s id="E2.1.1.4-1.2">In regions involved in agriculture, forest, and marine product related activities, the environmentally-conscious pursuit of these activities is helping maintain the environmental-protection capabilities of agricultural regions and forests.</s>
<s id="J2.1.1.4-1.2" corresp="E2.1.1.4-1.3 E2.1.1.4-1.4"> 一方、その生産活動に伴い、途上国を中心に進んでいる森林から農地への改変といった資源利用や欧米諸国における肥料等の使用による水質汚濁、家畜等からの温室効果ガスの一種であるメタンガスの発生という形で、環境に負荷を与えている。</s>	<s id="E2.1.1.4-1.3">On the other hand, as production activities are carried out, a load is placed on the environment because of changes in the intended use of resources.</s> <s id="E2.1.1.4-1.4">Examples of this include the conversion of forests to agricultural land, primarily in developing countries, the fouling of water from the use of fertilizers in Europe and North America, and the emission of methane gas (<cf><mol><a>C<a>H<n>4</mol></cf>), a type of greenhouse gas, by livestock.</s>
<div5 type=subsubsection id="J2.1.1.4.1">	<div5 type=subsubsection id="E2.1.1.4.1">
<head id="J2.1.1.4.1-h" corresp="E2.1.1.4.1-h"> ア．主食生産</head>	<head id="E2.1.1.4.1-h"> A. The production of staple foods</head>
<p id="J2.1.1.4.1-1" corresp="E2.1.1.4.1-1">	<p id="E2.1.1.4.1-1">
<s id="J2.1.1.4.1-1.1" corresp="E2.1.1.4.1-1.1 E2.1.1.4.1-1.2"> 世界の穀物生産量は、1965年の1006百万トンから1988年には1743百万トンと世界全体で73%増加し、特に、開発途上国では同期間に106%の大きな伸びを記録した。</s>	<s id="E2.1.1.4.1-1.1">World grain production rose from 1,006 million tons in 1965 to 1, 743 million tons in 1988, a 73% increase.</s> <s id="E2.1.1.4.1-1.2">In developing countries during this same period, there was a sharp 106% rise in grain production.</s>

Fig. 1 An Example from the bilingual corpus

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.

3. RWC text data

The Real World Computing (RWC) Program is a research program funded by the Japanese Ministry of International Trade and Industry (Mm) and it established the RWC Database Working Group in 1994 to gather and utilize real world knowledge. The Working Group is building several databases: text, speech, image and so on. Since 1994, the text group of the RWC Database Working Group has been building an annotated text database of modern Japanese for the research and evaluation of NLP technology [5].

The RWC text database should:

(1) be very large scale,
(2) include accurately annotated corpora, and
(3) be balanced and gathered from actual texts.

The following are the RWC text databases currently available.

(l)RWC-DB-TEXT-94-l

Morphologically analyzed data of MITI reports (manually post-edited; including MITI white papers for 1993-1995)

(2) RWC-DB-TEXT-94-2

Morphologically analyzed data of JEIDA’s annual report. (Manually post-edited. A survey report on the trend of natural language processing.)

(3)RWC-DB-TEXT-95-l

Differential data of the results of morphological analysis of the CD-Mainichi Shimbun (newspaper) (covers articles from the Mainichi Shimbun from 1991 to 1994).

This very large tagged corpus is the result of automatic morphological analysis of all sentences in CD-Mainichi Shimbun from 1991 to 1994. This 4-year database comprises about 100 million words (or morphemes).

Since RWC-DB-TEXT-95-1 is very large, it was only processed mechanically. Therefore it may include errors, although they are insignificant for practical use. We manually post-edited parts of RWC-DB-TEXT-95-1 to make RWC-DB-TEXT-95-2. This database will give basic data for estimating the accuracy of automatic tagging. Furthermore, RWC-DB-TEXT-95-2 can be used as training data for a morphological analyzer to make the remaining data in RWC-DB-TEXT-95-1 more accurate.

(4) RWC-DB-TEXT-95-2

Results of the morphological analysis of the CD-Mainichi Shimbun (Manually post-edited, 3,000 articles from 1994). This is the result of post-editing all sentences in the 3,000 articles extracted from RWC-DB-TEXT-95-1. These articles, the length of which ranged from 400 to 600 characters, were randomly selected.

(5) RWC-DB-TEXT-95-3

Articles tagged with the universal decimal classification (UDC). (30,000 articles from 1994.) This data is very useful for research on information retrieval.

We'll make following data available to the public.

(6) Iwanami Kokugo Jiten Data

Morphologically analyzed data from the Iwanami Shoten Japanese dictionary. (Manually post-edited)

Most of these data are tagged for Parts-of-Speech. As for syntactic tags, simple sentence data were extracted by the CRL from RWC-DB-TEXT-95-2 (described in the next section). Semantic tags will be added to Mainichi Shimbun data using Iwanami Kokugo Jiten.

4. Simple sentence data by the CRL and the IPA [6]

The Communications Research Laboratory (CRL) of the Japanese Ministry of Posts and Telecommunications extracted modifying relations between nouns and verbs (and adjectives) from real sentences in RWC-DB-TEXT-95-2 and made a simple sentence database from the extracted information. This data includes 59,939 simple sentences (with 3,243 different verbs) from 43,430 original sentences from newspaper articles. The Information Promotion Agency (IPA) also made the same type of database from its IPA corpus. This datum includes 108,262 simple sentences (with 5,161 different verbs) from 57,456 original sentences. Figure 2 shows an example of original sentences and two types of extracted simple sentences.

These simple sentence databases were developed focusing on changes of relations between particles in the original sentences and particles in the extracted simple sentences and describe the modifying relations between nouns and verbs. These are very useful (1) in surveying and making rules for actual linguistic phenomena, and (2) as training and/or reference data for NLP applications. CRL is doing research on improving an existing Japanese parser by using the frequency of patterns of modification which were extracted from the data and placed in a lexicon.

5. Conclusion

Several corpus projects, mainly JEIDA's bilingual corpus project, are explained in this paper. These corpora are being developed on the principles of being:

(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 develop public corpora for NLP research on these principles.

References

[1] Hitoshi Isahara: JEIDA's Test-Sets for Quality Evaluation of MT Systems -- Technical Evaluation from the Developer's Point of View --, MT Summit V, 1995.

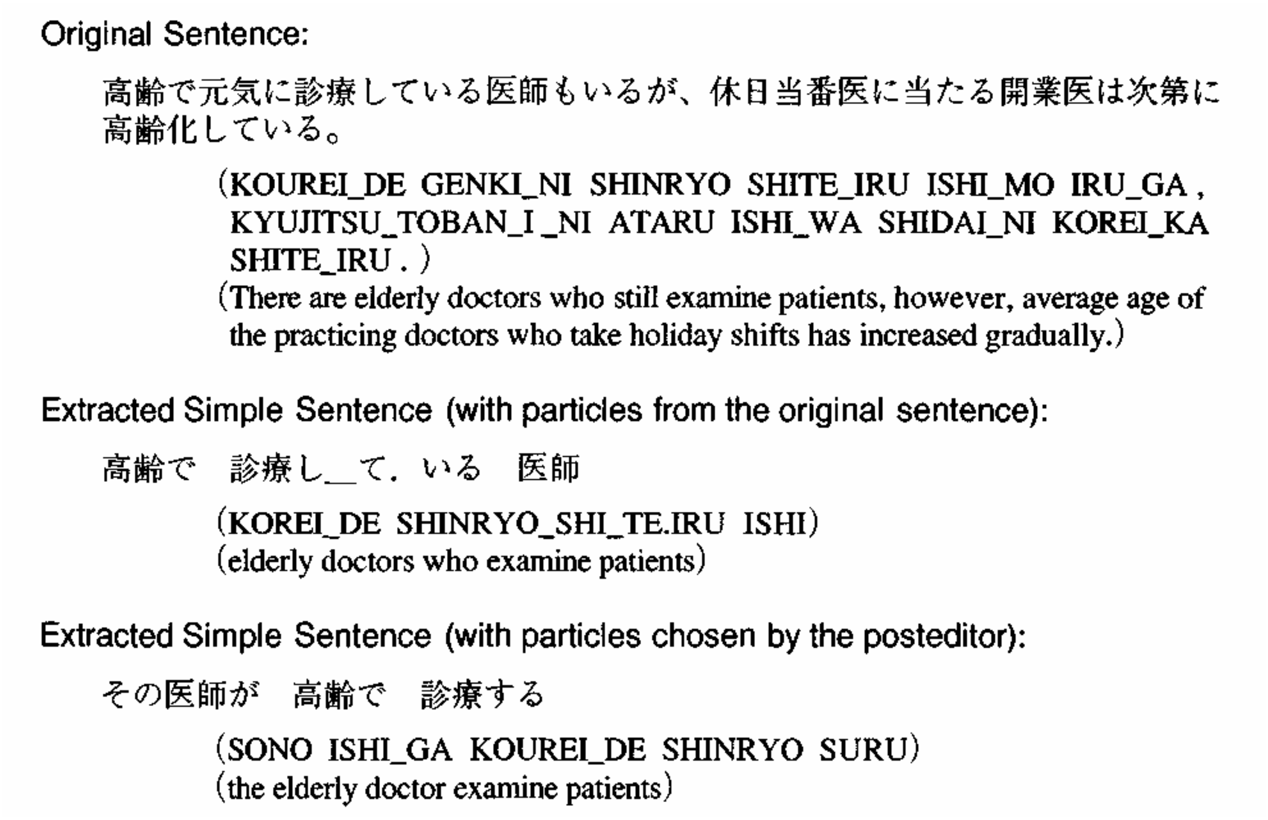
[2] Lou Burnard: TEI Lite: An Introduction to Text Encoding for Interchange, C. M. Sperberg-McQueen, 1995.

[3] Patrice Bonhomme et al.: LINGUA Information & Technical Aspect, Lingua Project, 1995.

[4] Eve Maler and Jeanne El Andaloussi : Developing SGML DTDs From Text to Model to Markup, Prentice Hall PTR, 1996.

[5] Hitoshi Isahara et al.: Building a text database with POS tags by RWCP (in Japanese), 1st Annual Meeting of Japanese Society of Natural Language Processing, 1995.

[6] Minako Hashimoto et al.: Extraction of simple sentence data from corpora (in Japanese), 3rd Annual Meeting of Japanese Society of Natural Language Processing, 1997.



Original Sentence:

高齢で元気に診療している医師もいるが、休日当番医に当たる開業医は次第に
高齢化している。

（KOUREI_DE GENKI_NI SHINRYO SHITE_IRU ISHI_MO IRU_GA ,
KYUJITSU_TOBAN_I _NI ATARU ISHI_WA SHIDAI_NI KOREI_KA
SHITE_IRU . ）

（There are elderly doctors who still examine patients, however, average age of the practicing doctors who take holiday shifts has increased gradually.）

Extracted Simple Sentence (with particles from the original sentence):

高齢で　診療し__て．いる　医師

（KOREI_DE SHINRYO_SHI_TE.IRU ISHI）

（elderly doctors who examine patients）

Extracted Simple Sentence (with particles chosen by the posteditor):

その医師が　高齢で　診療する

（SONO ISHI_GA KOUREI_DE SHINRYO SURU）

（the elderly doctor examine patients）

Fig. 2 An Example of the Simple Sentence Data