Still Tagging an Aligned Japanese/English Corpus

Francis Bond,† Yamato Takahashi,† Setsuo Yamada,† Makiko Nishigaki‡
†NTT Communication Science Laboratories
‡NTT Advanced Technology Corporation
bond@nttkb.ntt.jp

1 Introduction

This paper describes the continuation of work on tagging text used in research into Japanese-to-English machine translation, first described in (Bond *et al.* 1995).

As part of ongoing research into Japanese-to-English machine translation we have, for some time, been collecting various aligned parallel texts. The majority of these have been Japanese original texts for which we have prepared English translations. Creating texts in this manner has two advantages: (1) the translations are quite close to the original text, (2) they are aligned from the start. Unfortunately they also have a major disadvantage in that the translations are often stilted and unnatural, which limits their usefulness, particularly for any kind of statistical study. One major area where this kind of text is useful is in evaluation test-suites. Tagging such data will be discussed in Section 2.

Recently, we have been trying to gather aligned text where both the English and Japanese is of good quality. One such source is newspaper articles from databases (Shirai *et al.* 1995a; Shirai *et al.* 1995c; Shirai *et al.* 1995b). We discuss some of the issues of tagging such text in Section 3.

Finally, in Section 4, we discuss the explicit use of tags in machine translation, originally discussed in Bond *et al.* (1995).

1.1 Tag set used

The tag set used is a subset of the Text Encoding Initiative's P3 tag set (Sperberg-McQueen and Burnard 1994). The TEI tags were chosen for two reasons: because they offer a well though out framework in which to work; and because we hope that their increasingly wide adoption will make it easier to share data with other projects.

An example of a similar project using these tags is the **Lingua** Project, (Bonhomme and Romary 1995) a European project which aims at developing translation aids and language learning tools. They have gathered parallel texts in six European languages and marked the aligned sentences using the TEI P3 encoding. A multilingual concordance, and an explanation of the encoding system is available at:

`http://www.loria.fr/ bonhomme/lingua/`

2 Tagging a test-suite

One important use of parallel texts is in the evaluation of machine translation systems. The texts used for such evaluations tend to be associated with a great deal of meta-information, either in the form of explicit questions, as in the JEIDA test set (Isahara 1995), or comments as in the NTT test set (Ikehara *et al.* 1994).

The JEIDA test set marked the head of each sentence or comment with a regular code, allowing text to be searched using simple search engines such as `grep` or `awk`. This has the advantage that it is simple and robust, but the test set as a whole is not easy to examine, and it is easy for inconsistencies to creep in.

To make our test set available as both online text and in a visually pleasing printed format, we added an extra layer of abstraction to the NTT test set and encoded it with the TEI P3 tag set. In the future, we expect that with a powerful enough SGML browser, the tagged test set can be used as a hypertext document with no further editing. For the present we have created tools to output the basic Japanese/English sentence pairs as plain text, or text with mark up of the GENRE, DOMAIN and STRUCTURE suitable for input to the machine translation system **ALT-J/E**. As part of the extraction process, Japanese sentences are preprocessed into a form that can be translated by **ALT-J/E**. For example, half width characters are rewritten as full width characters, and △ or ▽ in front of numbers are converted to + or − respectively. We are in the process of creating a printed version of the test suite, created by running the master file through an SGML → LaTeX converter, which will show both the sentence pairs and their associated comments.

We give an example of a fully tagged example (one of the multiple possible senses of the verb 塗る *nuru* 'coat') in Figure 1. Note that the Japanese sentence has multiple possible English translations. We decided to include multiple translations in the test suite, with the one judged best listed first, because having multiple translations allows more flexibility in designing and evaluating MT systems. We even kept translations judged as poor and bad, both as examples of the limits of human translators and as guides as to what translations should be avoided. Notes and explanations to the examples are added using the `<note>` tag.

```
<div type="subsubsection">
   <head>「塗る」</head>
   <div id="0000" type="example">
        <p id="0001"><s id="0100" n="04050">私は机にニスを塗る。</s></p>
        <p lang="EN" id="0002">
            <s id="0200">I coat the desk with varnish.</s></p>
        <p lang="EN" id="0003">
            <s id="0300">I apply varnish to the desk.</s></p>
        <p lang="EN" id="0004">
            <s id="0400">I paint a desk with varnish.</s></p>
     <linkgrp type="alignment">
       <link resp="ALT_SGML ver0" targets="0100 0200" type="trans">
       <link resp="ALT_SGML ver0" targets="0100 0300" type="trans">
       <link resp="ALT_SGML ver0" targets="0100 0400" type="trans-poor">
     </linkgrp>
        <note target="0100">‘I varnish the desk.’ is also ok.</note>
  </div>
   ...
</div>
```

Figure 1: A test sentence with multiple translations.

3 Tagging semi-aligned text

In order to study the translation of newspaper articles, we are building up a corpus of Japanese and English text, mainly taken from the Nikkei Telecom Database. We found, however, that the articles are not so much translations of each other, but articles about the same subject in different languages. Although we can align text at the article level, it is not always possible to do so at the sentence level (Shirai *et al.* 1995b).

To construct the corpus we first download articles and attempt to align them automatically (Takahashi *et al.* 1996). Aligned articles are then automatically tagged using a perl script and stored as pairs of articles, as shown in Figure 2. All the information present in the original article is preserved, so that the original text can be recreated (with some difference in white space and line breaks) just by deleting all tags.

The aligned articles are then examined by humans, and sentences that correspond to each other are identified with `<link>` tags. At present, we only mark sentence-to-sentence correspondence. In many to one mappings, each pair is marked with a separate tag. To alleviate the drudgery of aligning sentences, we have created an extension to MULE's[1] `psgml-mode`[2] that hides the tags and highlights the headlines. This makes the articles easier to read. Sentence pairs can be linked by clicking on them.

The same tool that is used to extract sentence pairs from the test suite, described in Section 2, can also be used with the aligned newspaper articles, as they use the same basic tags.

4 Using markup in machine translation

Many of the possible uses of explicit markup discussed in Bond *et al.* (1995) have been implemented in the Japanese-to-English machine translation system **ALT-J/E** over the past year.

Use of explicit information

The use of meta-information that can be gained from tags has been expanded from just using information about GENRE, and DOMAIN, to exploiting structural markup. Nakaiwa *et al.* (1996) describes a method of using text structure, such as titles and headers, text divisions and positions within them, list headers, and related information to aid in the resolution of zero pronouns in technical documents. In addition structural information about text can be used to aid parsing, for example, section titles are often noun phrases rather than whole sentences,

Marking confidence in the translation

In Bond *et al.* (1995) we discussed the use of the *ana* attribute with the `<seg>` tag, to mark elements that the post editor should check with special care. This has been implemented for three cases: Unknown proper nouns (ana="un"), supplemented elements (ana="supp") and noun-triggered possessive pronouns (ana="trgpossp"). In addition to the full use of SGML tags, a short version was introduced for interactive use, an example is given in Sentence 1.

[1] Multi-lingual Extension to Emacs.
[2] Lennart Staflin's major mode for editing SGML documents.

```
<div type="align-hand">
<div id="4um0000" lang="MJ" decls="I.gen2 I.dom0511" type="article">
<head id="4um0001" type="main"> 日経 300 先物・大引け</head>
<opener><dateline id="4um0002"><date value="1995-07-25">95/07/25/15:39</date></dateline></opener>
<argument>
<list>
<label>Publication</label><item>NEWS</item>
<label>Id</label><item>##50725140</item>
<label>Genre</label><item>BB BQ</item>
<label>Length</label><item>125 字</item>
</list>
</argument>
<p id="4um0003"><s id="4um0100"> 大幅に反落。</s>
<s id="4um0200"> 前日比 5.9 ポイント安の 240.6 ポイントでこの日の取引を終えた。</s>
<s id="4um0300"> 日経平均先物が取引を終えた 15 時以降は、日経 300 先物で日経平均先物
の買いポジションをヘッジする動きが強まり、この日の安値で引けた。</s>
<s id="4um0400"> 売買高は 2700 枚と前日とほぼ同じ水準。</s></p>
</div>
<div id="6c80000" lang="EN" decls="I.gen2 I.dom0511" type="article">
<head id="6c80001">NIKKEI 300 FUTURES CLS: PLUNGE TOWARD CLOSING</head>
<opener><dateline id="6c80002"><date value="1995-07-25">95/07/25</date></dateline></opener>
<argument>
<list>
<label>Publication</label><item>NEWS</item>
<label>Source</label><item>NQN</item>
<label>Id</label><item>##50725140</item>
</list>
</argument>
<p id="6c80003"><s id="6c80100">SEPT. NIKKEI 300 FUTURES CONTRACT RETREATED SIGNIFICANTLY
TUESDAY, CLOSING DOWN 5.9 AT THE DAY'S LOW OF 240.6 YEN.</s>
<s id="6c80300">AFTER NIKKEI 225 FUTURES CLOSED, INVESTORS INCREASED SELLING TO HEDGE AGAINST
RISKS IN A FALL OF NIKKEI 225 FUTURES.</s>
<s id="6c80400">TURNOVER CAME TO 2,700 LOTS, SIMILAR TO YESTERDAY'S.</s></p>
</div>
    <linkgrp type="alignment">
    <link resp="yamato" targets="4um0000 6c80000">
            <link resp="fukui" targets="6c80100 4um0100">
            <link resp="fukui" targets="6c80300 4um0300">
            <link resp="fukui" targets="6c80400 4um0400">
            <link resp="fukui" targets="6c80100 4um0200">
    </linkgrp>
</div>
```

SGML encoding of two aligned articles

日経300先物・大引け
95/07/25/15:39
大幅に反落。
前日比 5.9 ポイント安の 240.6 ポイントでこの日の取引を終えた。
日経平均先物が取引を終えた 15 時以降は、日経 300 先物で日経平均先物の買いポジションをヘッジする動きが強まり、この日の安値で引けた。
売買高は 2700 枚と前日とほぼ同じ水準。

300 FUTURES CLS: PLUNGE TOWARD CLOSING
95/07/25
SEPT. NIKKEI 300 FUTURES CONTRACT RETREATED SIGNIFICANTLY TUESDAY, CLOSING DOWN 5.9 AT THE DAY'S LOW OF 240.6 YEN.
AFTER NIKKEI 225 FUTURES CLOSED, INVESTORS INCREASED SELLING TO HEDGE AGAINST RISKS IN A FALL OF NIKKEI 225 FUTURES.
TURNOVER CAME TO 2,700 LOTS, SIMILAR TO YESTERDAY'S.

Mule Display showing two aligned sentences

Figure 2: Two views of a pair of aligned newspaper articles

(1) 豪ボン社は、本部を移動した。

豪 ボン 社 は、 本部 を 移動 した。
gō-bon-company-TOP honbu-OBJ iten-shita
Australia's Bond Corp headquarter moved

'Australia's Bond Corp moved its headquarters.'

??? Australia Bon Corp ??? moved ??% its ??% headquarters.

Tagging output for evaluation

We also tag word parts of speech and phrase structures, using the TEI tags `<w>` for words and `<phr>` for phrases, although the output is almost unreadable without some kind of tool. At present, we only tag noun phrases[3], but other kinds of phrases, or indeed the whole sentence structure could also be tagged. As well as the full TEI tags, we have prepared short forms of the part-of-speech and phrase tags, both kinds of output are shown in Sentence 2.

(2) 象は鼻が長いが、豚は鼻が短い。

象 は 鼻 が 長い が、 豚 は 鼻 が
zou-TOP hana-SUB nagai ga, buta-TOP hana-SUB
elephant nose long but, pig nose
短い。
mijikai.
short.

Elephants have long trunks but pigs have short snouts.

```
<NP>Elephants&common-noun</NP> have&verb
long&adjective <NP>trunks&common-noun</NP>
but&word <NP>pigs&common-noun</NP>
have&verb short&adjective
<NP>snouts&common-noun</NP>.
```

```
<phr ana=NP TYPE=CO NUM=PL REF=GEN> <w
ana=common-noun>Elephants</w></phr> <w
ana=verb>have</w> <phr ana=NP TYPE=CO
NUM=PL REF=GEN> <w ana=adjective>long</w>
<w ana=common-noun>trunks</w> </phr> <w
ana=word>but</w> <phr ana=NP TYPE=CO NUM=PL
REF=GEN> <w ana=common-noun>pigs</w></phr>
<w ana=verb> have</w> <phr ana=NP TYPE=CO
NUM=PL REF=GEN> <w ana=adjective>short</w>
<w ana=common-noun>snouts</w> </phr>.
```

5 Conclusion

In this paper we describe work in progress tagging bilingual Japanese/English text. First we discuss some issues in tagging a test suite and a corpus of newspaper articles, and our own use of the TEI P3 tagset. Then we discuss the use of tags as aids in machine translation research. All the discussions are illustrated with examples.

[3]In addition to part-of-speech, we give the countability, number and referential use of the noun phrase.

References

Bond, Francis, Hiromi Nakaiwa, and Satoru Ikehara. 1995. Tagging an aligned Japanese/English corpus. In *1st Annual Meeting of the Association for Natural Language Processing*, 325–328. The Association for Natural Language Processing.

Bonhomme, Patrice, and Laurent Romary. 1995. The lingua parallel concordancing project: Managing multilingual texts for educational purpose. In *Language Engineering '95*.

Ikehara, Satoru, Satoshi Shirai, and Kentaro Ogura. 1994. Criteria for evaluating the linguistic quality of Japanese to English machine translations. *Journal of Japanese Society for Artificial Intelligence* 9.569–579. (in Japanese).

Isahara, Hitoshi. 1995. JEIDA's test-sets for quality evaluation of MT systems — technical evaluation from the developer's point of view —. In *The Fifth Machine Translation Summit: MT Summit V*.

Nakaiwa, Hiromi, Takahiro Uekado, Yayoi Nozawa, and Francis Bond. 1996. Resolving zero pronouns using textual structure. In *2nd Annual Meeting of the Association for Natural Language Processing*, 313–316. (in Japanese).

Shirai, Satoshi, Susumu Fujinami, Satoru Ikehara, Hiromi Ueda, and Hiroko Inoue. 1995a. Constructing an aligned Japanese/English corpus of newspaper articles (1) — basic structure and discussion —. In *Record of the 1995 Joint Conference of Electrical and Electronics Engineers in Kyushu*, p. 855. (in Japanese).

——, Mitsue Matsuo, Takako Seshimo, Susumu Fujinami, and Satoru Ikehara. 1995b. Constructing an aligned Japanese/English corpus of newspaper articles (3) — peculiarities of the articles and sentence alignment —. In *Record of the 1995 Joint Conference of Electrical and Electronics Engineers in Kyushu*, p. 857. (in Japanese).

——, Hiromi Ueda, Satsuki Abe, Susumu Fujinami, and Satoru Ikehara. 1995c. Constructing an aligned Japanese/English corpus of newspaper articles (2) — aligning articles taken from a database —. In *Record of the 1995 Joint Conference of Electrical and Electronics Engineers in Kyushu*, p. 856. (in Japanese).

Sperberg-McQueen, C. M., and Lou Burnard (eds.) 1994. *Guidelines for Electronic Text Encoding and Interchange*. Oxford: Chicago.

Takahashi, Yamato, Satoshi Shirai, Susumu Fujinami, Satoru Ikehara, Hiromi Ueda, and Hideyuki Matsushima. 1996. Automatically aligning newspaper articles from databases. In *2nd Annual Meeting of the Association for Natural Language Processing*, 201–204. (in Japanese).