Tagging an aligned Japanese/English Corpus

Francis Bond, Hiromi Nakaiwa, Satoru Ikehara
NTT Communication Science Laboratories
{bond,nakaiwa,ikehara}@nttkb.ntt.jp

1 Introduction

This paper describes work in progress on tagging text to be used in research into Japanese to English machine translation. We restrict our discussion to tagging the input and output of a Japanese to English machine translation system, the title of the paper is somewhat misleading, the text we discuss is only aligned in that it is the input and output of a machine translation system. 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.

2 Project description

Many collections of text have meta-information, outside of the raw text itself, that we would like to make available when translating it and also for research and/or evaluation.

We consider tagging text for three purposes: input to a machine translation system, post-editing automatically translated text and evaluation of machine translation output.

Each of these purposes calls for a different approach to tagging. For input to a machine translation system we should try to provide information that is likely to available for the texts the system is designed to translate. We consider three kinds of meta-information to be applicable: structure, genre, and domain, all of which can be described by the TEI core tag set.

For post editing of text the addition of some extra information can make the checker's job simpler. For example, if a system can mark words with a translations that are likely to be wrong, for example a noun judged to be a proper noun that did not appear in the dictionary, they can be listed and presented for checking automatically. Similarly if the system is not confident about a translation (the analysis had a low score or was under-specified) then it can be marked for special attention by the post editor. These properties can be represented using the TEI tags for analysis.

Evaluation for research and development requires identification and classification of problems as well as noting specific information about the problem such as it's relative seriousness or possible solutions. We describe how to use the <**note**> tag to encode this information.

2.1 Structure

Structural information about a document can be used to predict the style of text within it. For example section titles are often noun phrases rather than whole sentences, while sentences in lists often omit elements

that appear in the list headers. To successfully process these phenomena, information about the structure of a document needs to be available. As a minimum we consider information about the following structures to be necessary.

1. Lists & List headers
2. Text divisions & Titles/Headers
3. Paragraphs & Position within a paragraph
4. Sentences

The core TEI tag set gives tags for encoding all of these structures. Ideally structural encoding should be passed untouched through the system, so that the output keeps the same structure, and, at the same time, the encoding should be available as meta-information for the translation engine itself. To do this, it is necessary to pass meta-information to the translation system with each sentence. For example by encoding its position within a text, its structural properties (TITLE, LIST HEADER. LIST ELEMENT, PARAGRAPH LEAD SENTENCE ...) and the relevant domain and genre. The meta-information can be provided with each sentence by an sgml parser.

2.2 Genre and Domain

To label the genre and domain of a text or text fragment we use the <**textclass**> tag. This allows us to define the scheme we use in the header, and refer to it as an attribute to units of text from the text itself down to a single paragraph. In practice we distinguish between eight genres and 99 different domains, using an in-house classification scheme. It is easy to define another classification scheme in the header and use both at once.

We define our classification scheme in the <**encodingdescription**> and make it available to be used to mark chunks of text as a <**textclass**> as shown in Figure 1. It can now be used to mark, for example, a paragraph as being about sport as follows: <**p decls="I.dom02"**> Kendo is the Japanese art of swordsmanship ... <**\p**>.

2.3 Marking confidence in the translation.

For tagging the output of a machine language system, we use the multipurpose tag <**seg**> along with the attribute *ana*. <**seg**> is used to mark an arbitrary segment of text, and *ana* is used to point it to an interpretation. The possible interpretations can be listed as a collection of <**interp**>s within an <**interpgrp**>. An example of an <**interpgrp**> is given in Figure 2.

Segments in the output can then be marked with these interpretations. Consider the following sentence with the machine translation given.

(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。
Australia-bon-company-TOP	headquarters-OBJ	moved.
Australia's Bond Corp	headquarters	moved.

```
<teiheader>
  <filedesc> ...  </filedesc>
  <encodingdesc>
    <classdecl>
      <taxonomy id=I>
        <category id =I.gen><catdesc>genre</catdesc>
          <category id =I.gen0><catdesc>General</catdesc> </category>
          <category id =I.gen1><catdesc>Newspaper</catdesc> </category>
          <category id =I.gen2><catdesc>Technical manual</catdesc> </category>
          <category id =I.gen3><catdesc>Academic paper</catdesc> </category>
        </category>
        <category id =I.dom><catdesc>domain</catdesc>
          <category id =I.dom00><catdesc>General</catdesc> </category>
          <category id =I.dom01><catdesc>Mathematics</catdesc> </category>
          <category id =I.dom02><catdesc>Sport</catdesc> </category>
          <category id =I.dom03><catdesc>Food</catdesc> </category>
        </category>
      </taxonomy>
    </classdecl>
  </encodingdesc>
  <profiledesc>
    <textclass>
      <catref target="I.gen I.dom">
    </textclass>
  </profiledesc>
 </teiheader>
```

図 1: Header declarations for genre and domain

```
   <text>
    <interpgrp resp="system">
      <interp id=possp type="MT"  value="Default possessive pronoun">
      <interp id=un    type="MT"  value="Unknown compound noun">
    </interpgrp>
    <body> ... </body>
   </text>
```

図 2: Interpretations of machine translation text.

'Australia's Bond Corp moved its headquarters.'

Australia Bon Corp moved its headquarters.

The MT output can be automatically marked as: '<**seg** ana=un> Australia Bon Corp <**\seg**> moved <**seg** ana=possp> its <**\seg**> headquarters.' The marked segments would then be highlighted in some way for the post editor, for example using a different font to give: 'AUSTRALIA BON CORP moved *its* headquarters.' This makes it simpler to check for problems. In this case the unknown noun phrase needs to be corrected, but the possessive pronoun can remain.

2.4 Tagging output for evaluation.

We use <**seg**> and *ana*, along with the general purpose <**note**> as aids for evaluation of MT output. Problems are identified with <**seg**> and classified either using *ana* to point to a problem class which needs no further explanation or to a note, in which a specific problem can be addressed. The <**note**> tag includes the attributes *resp* to specify who is responsible for the note, *type* used to identify the problem type and *target* which identifies to what the note applies.

The notes can be kept seperate from the translated text as they are linked with the target reference and can therefore be added and deleted at will.

3 Conclusions

We have described some uses of TEI tags as aids in machine translation research. They can provide information used by the translation engine itself, as well encode information from the system that is useful for post editors. Finally the laborious process of evaluating machine translation systems can be made simpler by using tags to identify and classify problems.

Tools and Data

The tags described in this paper are from the TEI 3 core encoding with the analysis extension. The document type definitions are available by anonymous ftp from ftp.ifi.uio.no in directory TEI.

参考文献

SPERBERG-MCQUEEN, C. M., and LOU BURNARD (eds.) 1994. *Guidelines for Electronic Text Encoding and Interchange*. Oxford: Chicago.