Alignment Support for Multi-lingual humanity Corpus Retrieval
Case Study: The Little Prince and the Tale of Genji

Hiroko Omae†, Frederic Andres††, Michita Imai†††and Yuichiro Anzai†††
†Graduate School of Science and Technology, Keio University
hiroko@nii.ac.jp
††National Institute of Informatics
†††Faculty of Science and Technology, Keio University

In this paper, we propose an alignment support tool for multi-lingual humanity corpus retrieval. We show that alignment processing of multi-lingual documents helps the access to multi-lingual contents. The System creates a tree-structure by analyzing linguistic tags from input documents and defines the alignment by matching the keyword. It visualizes the tree-structure and alignment of each segment. It also makes the glossary semi-automatically, which we can use during the translation process. Furthermore, experimentation of the prototype using "The Little Prince" of Antoine de Saint-Exupery and "The Tale of Genji" (contemporary style version) of Murasaki-Shikibu are shown. We evaluated the system by making the alignment of Japanese-French/Japanese-English documents.

1. Introduction

The activity to digitalize and to conserve a cultural heritage is growing, also the need of digital data distributing is increasing. Moreover, since a classical literature is often used to introduce a specific culture to other countries, such content will be digitalized and distributed. So translating digital documents becomes a very important task.

In the field of translation, the research on machine translation has been done during the past 20 years. However, recently, it is said that the translation made by machine alone has a limit on its accuracy. Therefore, the machine aided translation system, which allows human to work together, is noticed. Especially, the translation memory is the function to realize the example-based translation system and it is said that this function is available to the translation of the specific domain.

As for the translation of a classical literature, it also seems to be available to use the translation memory, but there is a difficulty of gaining the correspondence of the words since the terms are sometimes so old that we cannot find them in the dictionaries. In this paper, we propose the alignment support tool that support finding the correspondence of the sentences and words (called "alignment") and to generate the glossary semi-automatically. This tool enables to define the correspondence of unknown words easily and to apply the glossary generated by this tool to the translation using translation memory.

Users can modify the correspondence when the result shown by the system is incorrect, and thus, this tool avoids making the wrong translation by finding the wrong correspondence and generating wrong glossary. We used the Japanese, French and English documents, such as "Little Prince" of Antoine de Saint-Exupery and "The Tale of Genji" (contemporary style version) of Murasaki-Shikibu.

This paper is organized as follows. Section 2 describes related works. Then Section 3 describes the "alignment" tool in details. In Section 4, we show the experiment, the result, and the evaluation. We discuss about future work in Section 5 and we conclude in Section 6.

2. Related Works

2.1 Translation Memory

Translation Memory[1)] represents the function to store the source-target sentence pair, which was translated once, in the database as translation pattern. When the new sentence is input, the example-based translation system searches the pattern that matches or resembles it and outputs searched pattern by modifying it if necessary.

Using translation memory enables to decrease the translation cost, and it is said to be available to translate the documents that need few modification, such as revising the manuals of upgraded software. Besides, if the sentence that match input sentence does not exist in the translation memory, system can store this sentence as translation pattern to increase the accuracy. Therefore, it is very convenient when we make much translation in the same domain.

Figure 1 shows the system using the translation memory.

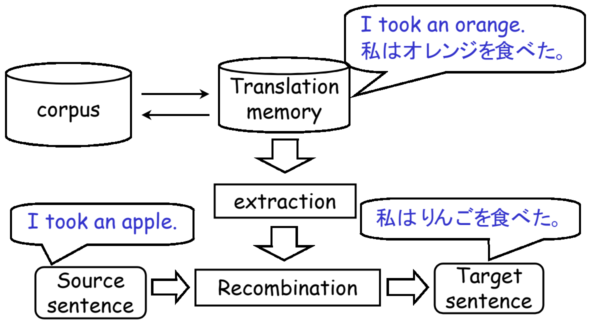


Fig.1 system using translation memory

2.2 Machine Aided Translation Tools

In this section, we introduce two machine aided translation tools that are related to our research and discuss their problems.

- Waligner
  Waligner[3)] is a Thai-English document alignment tool. At first, it extracts the corresponding sentences of each document using common known-words and their distance, then makes the correspondence of unknown-words. It suggests several words as candidates and user can choose the correct word. However, there are cases that many candidates are suggested but there is no correct answer.
- JGloss
  JGloss[4)] is a Japanese document annotation support tool. When the Japanese document is input, the system analyses it morphologically and shows the reading of Kanji and English translation of each word. If the punctuation, reading, or translation of the phrase is wrong, user can revise it and generate a user dictionary. However, since this tool is not an alignment support tool, it cannot make the correspondence between two language documents. In addition, it shows the English translation of each word but not the translation of whole sentence. Furthermore, there is also another problem. As this system needs to use the specific tools such as the Japanese-English dictionary called EDICT or the morphological analyzer called Chasen, its accuracy depends on these tools.

2.3 Linguistic DS

Linguistic DS[2)] is an XML tag set of GDA (Global Document Annotation) focusing to annotate multi-lingual documents syntactically and semantically. It is standardized by Mpeg-7. Linguistic DS enables to create a semantic structure of the document and to describe multi-lingual documents in unified format.

Table 1 shows the example of document tagged in Linguistic DS.

Table 1 Example of Linguistic DS

In the beginning God created the heaven and the earth.

```
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001" xml:lang="en">
 <Description xsi:type="ContentEntityType">
  <MultimediaContent xsi:type="LinguisticType">
   <Linguistic>
    <Sentence id="b.GEN.1.1" type="verse">
     <Phrase>In
      <Phrase>
       <Phrase>the </Phrase>
       beginning
      </Phrase>
     </Phrase>
     <Phrase id="GOD">God </Phrase>
     created
     <Phrase>
      <Phrase id="HEAVEN">
       <Phrase>the </Phrase>
       heaven
      </Phrase>
      and
      <Phrase id="EARTH">
       <Phrase>the </Phrase>
       earth
      </Phrase>
     </Phrase>
     .
    </Sentence>
   </Linguistic>
  </MultimediaContent>
 </Description>
</Mpeg7>
```

By using Linguistic DS, we can describe the modification and verbal dependency relation or pronoun and antecedent relation easily. In this paper, we handle the document tagged simply as it is shown in table 1.

3. Alignment support tool

3.1 System Overview

The overview of the system is shown in Figure 2.

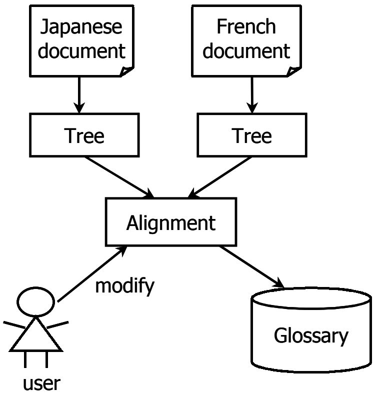


Fig.2 System Overview

The documents that are input to the system have to be tagged in Linguistic DS. The system shows these documents as tree structure by using tags. It also finds and shows the correspondence of words automatically by using attributes. If the result shown by the system is wrong, user can modify it, then, store it in the glossary.

3.2 Sentences Alignment and Tree Creation

The system creates the tree structure from the document tagged in Linguistic DS. When the tagged-document file is input to the system, it shows the plain text. If the user chooses one sentence in this text, it shows the tree structure of selected sentence and finds sentence correspondence in the other document. Correspondences can be made from both languages.

3.3 Word Alignment

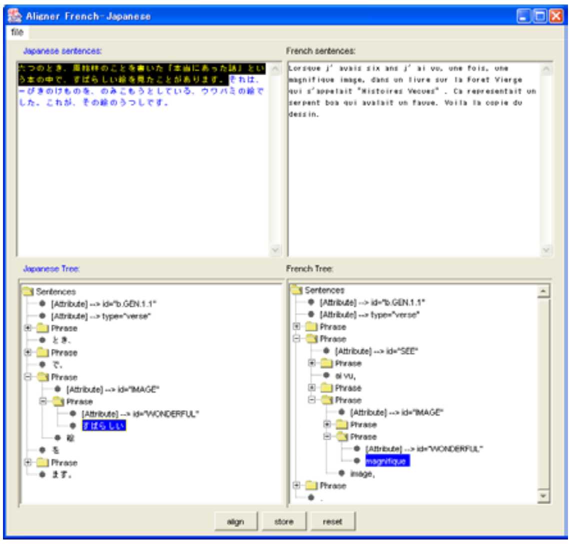
System generates the word alignment from the tree structure. Also in this case, the correspondence can be made from both languages. In Linguistic DS, the important words are tagged with a keyword as an attribute. The system finds the correspondence using this keyword and visualizes its relation. To handle the words which the keywords are not given, the system uses the other attributes. If the result shown by the system is not correct, the user can modify the result.

3.4 Glossary Generation

After modifying the result, the user can register the words with the dictionary. To allow the search from both sides, the system creates two dictionaries for each language. Words are simply sorted by ASCII code order.

4. Examination Example

In this paper, we made an experiment using “The Small Prince” of Antoine de Saint-Exupery and the “Tale of Genji” (contemporary style version) of Murasaki-Shikibu as an example. This tool is implemented in J2SDK1.4.1 Figure 3 is the snapshot.



Japanese sentences:

六つのとき、原始林のことを書いた「本当にあった話」という本の中で、すばらしい絵を見たことがあります。それは、一ぴきのけものを、のみこもうとしている、ウワバミの絵でした。これが、その絵のうつしです。

French sentences:

Lorsque j' avais six ans j' ai vu, une fois, une magnifique image, dans un livre sur la Foret Vierge qui s'appelait "Histoires Vecues" . Ca representait un serpent boa qui avalait un fauve. Voila la copie du dessin.

Japanese Tree:

```
Sentences
    [Attribute] --> id="b.GEN.1.1"
    [Attribute] --> type="verse"
    Phrase
    とき、
    Phrase
    で、
    Phrase
        [Attribute] --> id="IMAGE"
        Phrase
            [Attribute] --> id="WONDERFUL"
            すばらしい
        絵
    を
    Phrase
    ます。
```

French Tree:

```
Sentences
    [Attribute] --> id="b.GEN.1.1"
    [Attribute] --> type="verse"
    Phrase
    Phrase
        [Attribute] --> id="SEE"
        Phrase
        ai vu,
        Phrase
        Phrase
            [Attribute] --> id="IMAGE"
            Phrase
            Phrase
                [Attribute] --> id="WONDERFUL"
                magnifique
            image,
    Phrase
    .
```

Fig.3 Examination example

We examined the first sentence of the first chapter in "The Little Prince". Below are the French and Japanese sentences.

Lorsque j'avais six ans j'ai vu, une fois, une magnifique image, dans un livre sur la Foret Vierge qui s'appelait "Histoires Vecues".
六つのとき、原始林のことを書いた「本当にあった話」という本の中で、すばらしい絵を見たことがあります。

In this example, we could gain and store "magnifique—すばらしい", "image—絵", "livre—本", "Foret Vierge—原始林".

In French, there are conjugations of verbs, nouns and adjectives. In this case, the system could not gain and store the original form, but by giving the original form of each part of speech, as attribute, it will solve it, and it can easily be realized.

We also examined the first sentence of the first chapter in "Kiritsubo" of "The Tale of Genji". Japanese and English sentences are as follows:

どの帝の御代のことであったか、女御や更衣たちが大勢お仕えなさっていたなかに、たいして高貴な身分ではないで、きわだってご寵愛をあつめていらっしゃる方があった。
At the Court of an Emperor there was among the many gentlewomen of the Wardrobe and Chamber one, who though she was not of very high rank was favoured far beyond all the rest.

In this sentence, we could get and store "帝—Emperor", "御代—Court", "身分—rank".

However, in this case, the word "女御" is translated as "gentlewomen of the Wardrobe" and the word "更衣", as "gentlewomen of the Chamber". The system cannot recognize the correspondence when more than two words correspond to one word like this example, so that the user have to specify the corresponding words manually. We also intend to improve the system in this point.

5. Future Work

Linguistic DS is a tag set that can handle the semantic structure and enables it to be represented in detail. Currently, our system can handle only the documents that are simply tagged with few types of tags, but it is possible to increase the accuracy and to reduce the user's task by using the documents tagged in detail.

Besides, tagging is the task that costs the most, and it is very difficult to develop an automatic tagging system. Therefore, we aim to create an annotation editor of the Linguistic DS to reduce the tagging cost. Also, it is our future task to make a quantitative evaluation of this tool by tagging a long composition using above editor.

6. Conclusion

In this paper, we proposed the alignment support tool for multi-lingual humanity corpus retrieval. This tool enables to generate the glossary, which is required for translation, semi-automatically and enables users to modify its result. We have implemented the system and executed it using contemporary version of "The Tale of Genji" of Murasaki-Shikibu and "The Little Prince" of Antoine de Saint-Exupery. This system will be extended in several points to be integrated into machine aided translation system.

Reference

[1] Emmanuel Planas, Osamu Furuse: "Formalizing Translation Memories," Proceedings of Machine Translation Summit VII, Singapore, 1999.

[2] Hasida K., Andres F.,Boitet C., Calzolari N., Declerck T., Farshad Fotouhi, William Grosky, Shun Ishizaki, Asanee Kawtrakul, Mathieu Lafourcade, Katashi Nagao, Hammam Riza, Virach Sornlertlamvanich, Remi Zajac, Zampolli, A.: "Linguistic DS," IO/IEC JTC1/SC29/WG11, MPEG2001/ M7818

[3] Nithiwat Kampanya, Prachya Boonkwan, Asanee Kawtrakul: "Bilingual Unknown Word Alignment Tool for English-Thai," Joint International Conference of SNLP-Oriental COCOSDA 2002, 9-11 May 2002, Hua Hin, Prachuapkirikhan, Thailand

[4] Michael Koch: "JGloss User's Guide," 2002.

[5] Van Zaanen M.: "ABL: Alignment-Based Learning," Proceeding of COLING2000, International Conference on Computational Linguistics, Saarbruecken, Germany, 2000.

[6] Murasaki Shikibu, translated by Arthur Waley: "The Tale Of Genji," Tuttle Publishing, 2002.