Word Alignment Viewer for Long Sentences

Hideki Kashioka
Spoken Language Communication Research Labs., ATR
2-2-2 Hikaridai "Keihanna Science City", Kyoto,
619-0288
Japan,
hideki.kashioka@atr.jp

Abstract

An aligned corpus is an important resource for developing machine translation systems. We consider suitable units for constructing the translation model through observing an aligned parallel corpus. We examine the characteristics of the aligned corpus. Long sentences are especially difficult for word alignment because the sentences can become very complicated. Also, each (source/target) word has a higher possibility to correspond to the (target/source) word. This paper introduces an alignment viewer a developer can use to correct alignment information. We discuss using the viewer on a patent parallel corpus because sentences in patents are often long and complicated.

1 Introduction

Recently, many researchers have focused their interest on statistical machine translation (SMT) systems, with particular attention given to models (including translation units) and/or decoding algorithms. There are many kinds of translation units, i.e., word, n-gram, phrase, clause and so on, based on various proposed translation models. These translation units are used not only in SMT systems but also other methods' MT. The aligned data are important resources when the developer of am MT system wants to select the translation unit for translation model; the aligned data would be indicate the most suitable unit for translation model. Developers carefully examine the aligned data and considers the characteristics of those data in order to choose a suitable unit for their translation models. A well designed viewer for displaying aligned data would be useful in such a situation. When we look at the aligned data, we actually need to view some different points to make sense of the information. In some cases, we want to look at full sentences and overviews for each piece of word aligned information. In another cases, we want to see whether some part of a sentence and local information are important. To do this, the viewer needs some display modes, depending on the purpose. In this paper, we illustrate our developed viewer for displaying/editing the aligned data.

2 Required Functions of Viewer

In this section, we describe the functions that are required in the aligned-data viewer. The following three points are required functions when a developer is considering a translation model.

2.1 Function for graphical illustration

The graphical illustration of the alignment information is a natural and basic function. We should focus on which information is important and how to effectively illustrate important information for a translation model.

There are at least two types of graphical illustration of the alignment information for long sentences. First, full sentence is shown and considering with full sentence. This kind of information tells us the types of word-position exchanging information present in the sentence, gives an overview of the components with syntactical structure information, and so on. Second, some parts of sentences are shown in consideration to detailed local parts of sentences. This kind of information provides us with data on compound words, idiomatic phrase mapping, and so on.

2.2 Function for editing alignment information

Our current belief is that word alignment information sometimes includes errors and/or includes other possible or suitable candidates for word alignment for the translation model. When we find such errors or other candidates during consideration of the translation unit, we want to correct/modify aligned data. Furthermore, when we make test sets for word alignment, we need a function for editing alignment

information.

2.3 Function for segmenting the sample unit

When the viewer displays the alignment information as it is, we need a function to illustrate the segmentation that makes pseudo-units for the translation model easy to check. Combining this function with the function for graphical illustration would be beneficial.

3 Viewer Design

In this section, we illustrate the implemented viewer and the environment information.

3.1 Treated corpus

Our viewer under development does not depend on the domain of the target data. In this paper, we explain our viewer using a patent parallel corpus comprising patents made in Japan in Japanese and patent abstracts translated from Japanese to English that were published in 2003.

The alignment information was not included in the original patent data. We obtained the alignment information using GIZA++(Och and Ney, 2000), which is an extension of the program GIZA for learning statistical translation models from bitext. GIZA++ outputs the alignment file and the translation table. We use these two files for our viewer to display alignment information.

In the alignment file, the best (Vitterbi) alignment is written on three lines for each sentence pair: the first line is a label, the second line is the target sentence and the third line is the source sentence. Alignment information is represented in the source sentence. Each token is followed by a set of numbers that indicate the positions of the target words to which the source word is aligned. The "NULL" token is added to the source sentence. This token is a special one that illustrates the set of target words that have no connection to the source words.

The word segmentation of Japanese sentences is produced by the Japanese morphological analyzer "ChaSen"(Asahara and Matsumoto, 2000).

3.2 Implementation of the viewer

Our developed viewer[1] has three modes. Two modes are for displaying the alignment information and one other one is for showing structural information on each sentence. The first mode is shown in Fig.1. This mode has three main areas. The upper-right area in this mode displays source and target sentences, and the alignment information of each segment is shown in the same color for both source and target sentences. When the mouse cursor placed over a word, more detailed information can be shown by highlighting the corresponding word, and the positions of highlighted words are shown at the bottom of this area. The lower-right area in this window represents the word alignment information according to the position of the source and the target word in each sentence. The text in this area is editable, so we can change the alignment information by editing this area's information. Pressing the "F5" key provides a representation of the edited information in this area. In Fig.1, the words "コスト" and "cost" are highlighted and they illustrate the position information as "Sentence=PBS0 Japanese=j24 English=e49" in the center-right area. This position information is presented on the text in lower-right area in this window. If the text is edited, the highlighting and the position information in the center-right area will also change. The left-hand side of this window shows a file included in the parallel corpus.

The second mode is shown in Fig.2. The left-hand area in this mode looks the same as that in Fig.1, while the right-hand shows a grid representation with the X-axis indicating the source language and the Y-axis showing the target language. The connection between source and target words is shown by solid circles. The number printed to the right-hand side of each circle is the probability of the pair's successful translation. In this mode, the alignment information can be edited by clicking the grid. Also in this mode, the expression of English (Y-axis) is not by just one word: if the Y-axis were just one word, it would be too long and we would need an enormous area to display the grid.

These English expressions are defined by a parse tree[2]. When the words represented by leaf nodes in the parse tree have a relationship with brother nodes, they form an expression[3]. In addition, some glid lines are thick. In Japanese, these thick lines (bars) indicate borders between clauses. Clause boundaries are annotated in preprocessing. The crossed thick lines indicate

[1]This viewer was implemented via JAVA on Windows.

[2]Now we use the Charniak Parser(Charniak, 2000)

[3]Japanese (X-axis) do not process with the concatenation.

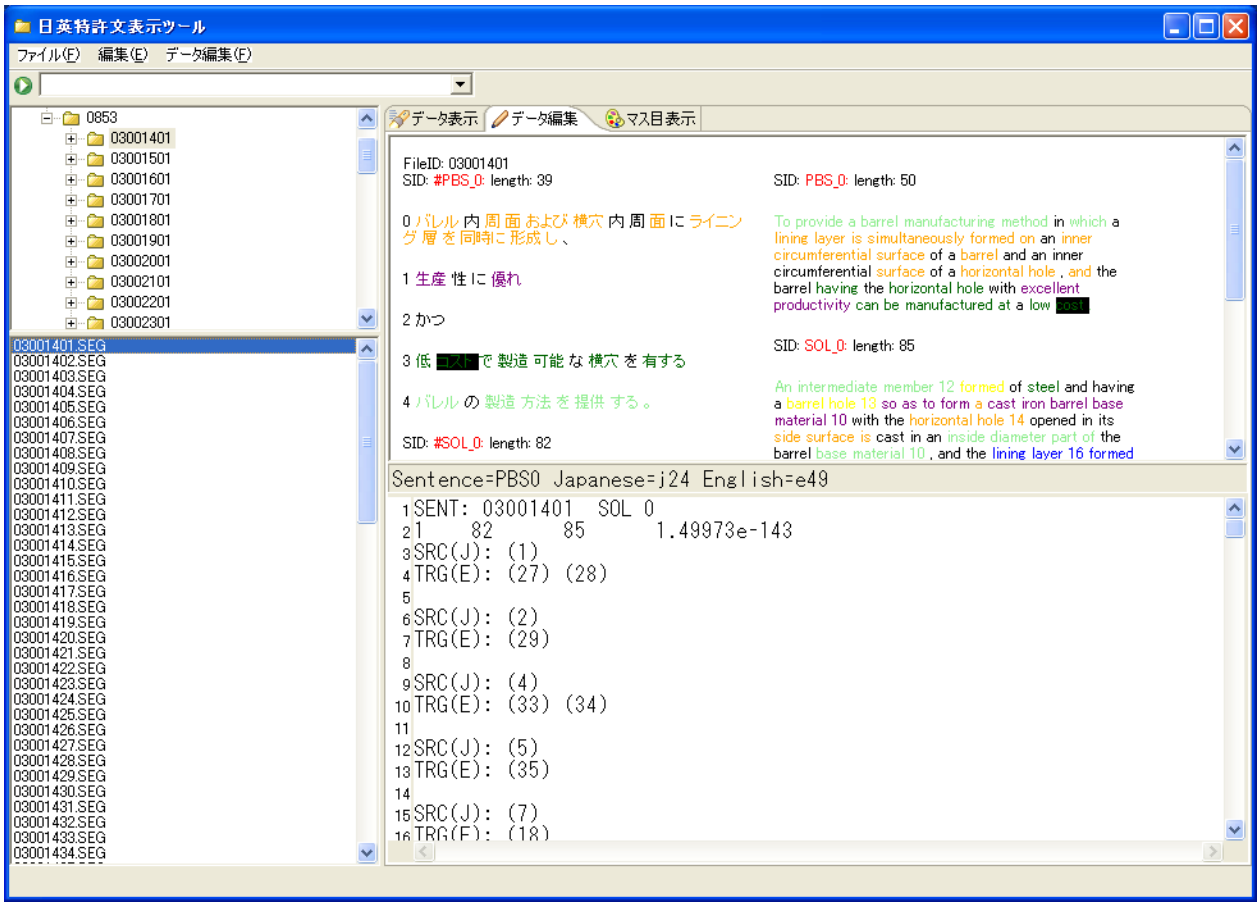

Figure 1: Screenshot of the word alignment viewer with full sentences

the border of a constituent with its corresponding Japanese clause. There are several ways to calculate the borders of the corresponding Japanese clauses. The processing will be discussed in detail in the next section.

These thick lines help to determine whether the segmentations are suitable for the translation model.

The third mode is shown in Fig.3. This mode mainly illustrates the analysis result for each sentence. The left-hand area in this mode is the same as in the other modes. The lower-middle area in this mode shows morphological analysis and clause-boundary annotation results for the Japanese by ChaSen and CBAP(Kashioka et al., 2003). The lower-right area shows the parse result with English[4], while the upper-right area shows the English constituent list that was segmented with corresponding Japanese clauses.

[4]Currently we use the Charniak Parser to get these results.

4 Discussion

In this section, we would like to discuss alignment information. The patent parallel corpus includes long sentences and these are difficult to align automatically. In the previous section, we used the alignment information automatically calculated by GIZA++. This viewer can display pseudo-units as Japanese clauses for translation. We would like to construct a model based on Japanese clauses as translation units, and we are considering incorporating this viewer into the model. Unfortunately, the word segmentation on the Japanese side and alignment information are poor because patent sentences include a lot of technical terms and each sentence is long. Thus we need to obtain more accurate data. We are considering an alignment method that modifies GIZA's output by using Japanese clause segment information. When examining this method, this viewer will be helpful.

Currently three methods are implemented as the processings for segmentation. The first is a very simple segmentation, which only checks

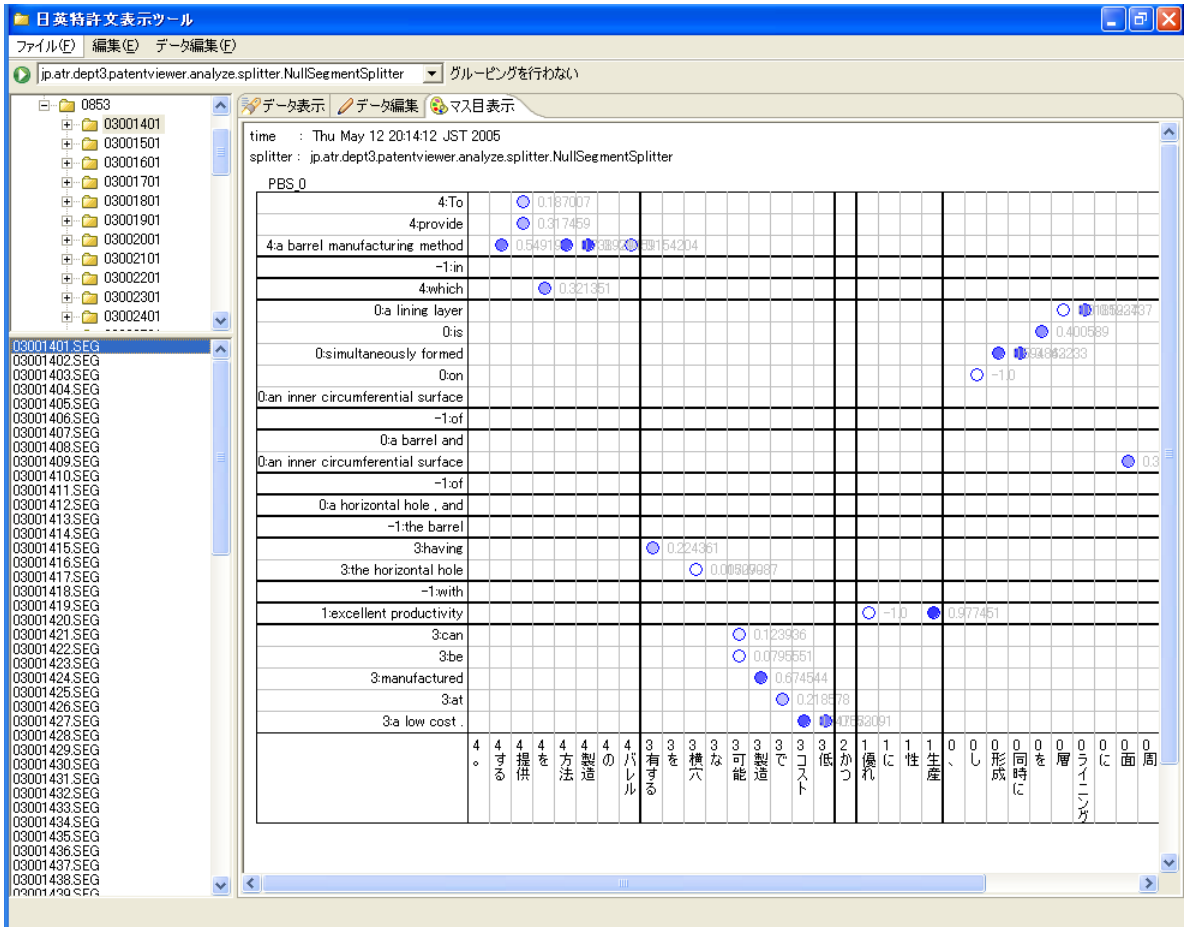

Figure 2: Screenshot of the word alignment viewer with parts of sentences

whether a word connected to the following word in English belongs to the same Japanese clause. When these words do not belong to the same clause, it creates a segment position.

The second is a segmentation that adds a null connection process. When an English word has no connected word, that English word is concatenated with the previous segment.

The third is a segmentation using a translation probability. Concatenate processing is performed with words that have higher translation probability values.

These segmentation results are slightly different due to different processes. We are considering a re-alignment method to obtain more accurate results using this information.

5 Conclusion

This paper introduced our viewer under development, which has three modes for displaying a Japanese-English patent parallel corpus. The alignment information uses GIZA's output. Using this viewer, the alignment information can be corrected manually. We could observe alignment information with full and/or partial sentences. We are now planning to produce correct alignment data for MT evaluation, and are trying to construct a translation model based on clauses as translation units, using our viewer to check this unit.

References

Masayuki Asahara and Yuji Matsumoto. 2000. Extended models and tools for high-performance part-of-speech tagger. In *Proceeding of COLING 2000*, Saarbrucken, July.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of NAACL 2000*, pages 132–139, April.

Hideki Kashioka, Takehiko Maruyama, and Hideki Tanaka. 2003. Building a parallel corpus for monologue with clause alignment. In *Proceeding of MT Summit IX*, pages 216–223, September.

F. J. Och and H. Ney. 2000. Improved statisti-

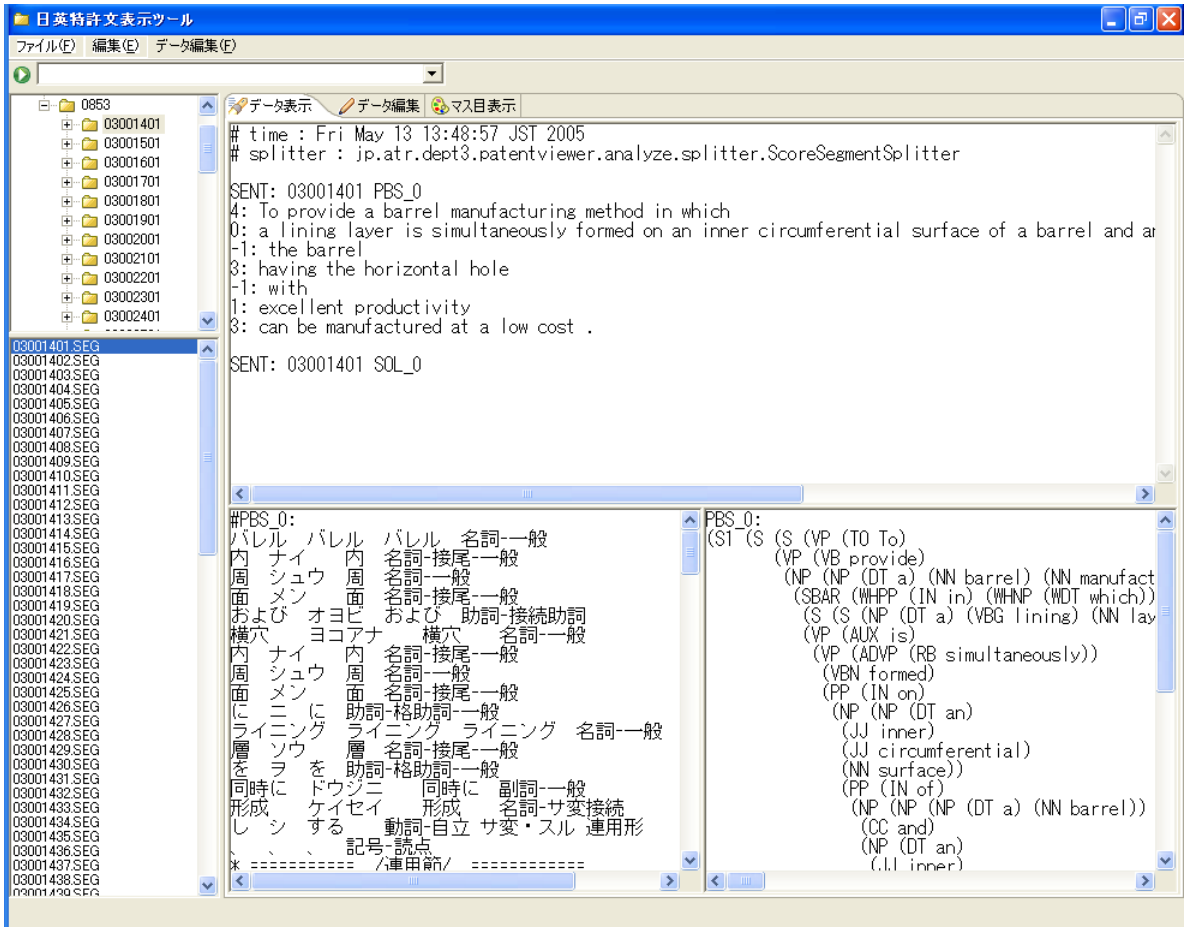

Figure 3: Screenshot of the word alignment viewer for sentence analysis results

cal alignment models. In *Proceeding of ACL 2000*, pages 440–447, Hongkong, China, October.