Description of KYOTO EBMT System in PatentMT at NTCIR-10

Toshiaki Nakazawa Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, 606-8501, Japan
{nakazawa, kuro}@nlp.ist.i.kyoto-u.ac.jp

ABSTRACT

This paper describes "KYOTO" EBMT system that attended PatentMT at NTCIR-10. When translating very different language pairs such as Japanese-English, it is very important to handle sentences in tree structures to overcome the difference. Many of recent studies incorporate tree structures in some parts of translation process, but not all the way from model training (parallel sentence alignment) to decoding. "KYOTO" system is a fully tree-based translation system where we use the treelet alignment model by bilingual generation and monolingual derivation on dependency trees. and example-based translation.

Team Name

KYOTO

Subtasks

Japanese to English, English to Japanese

Keywords

Fully Syntactic EBMT, Word Dependency Tree, Bilingual Generation and Monolingual Derivation

1. INTRODUCTION

We consider that it is quite important to use linguistic information in translation process when tackling on very different language pairs such as Japanese and English, and one of the most important information is a sentence structure. Many of recent studies incorporate some structural information into decoding, rarely into alignment. In this paper, we propose a fully tree-based translation framework based on dependency tree structures. In the alignment, we use the treelet alignment model by bilingual generation and monolingual derivation based on dependency trees [11]. Section 2 shows a brief description of the model. It is a kind of tree-based reordering model, and can capture non-local reorderings which sequential word-based models cannot often handle properly. Furthermore, the monolingual derivation model can capture the difference of the function word set of two languages.

In the translation, we adopt an example-based machine translation (EBMT) system [10], handling examples which are discontinuous as a word sequence, but continuous structurally. It also considers similarities of neighboring nodes, which is useful for choosing suitable examples matching the context.

Figure 1 shows the overview of our EBMT system on Japanese-English translation. The translation example database is automatically constructed from training parallel corpus by means of treelet alignment model. Note that both source and target sides of all the examples are stored in dependency tree structures. An input sentence is also parsed and transformed into dependency structure. For all the treelets in the input dependency structure, matching examples are searched in the example database. This step is the most time consuming part, and we exploit a fast tree retrieval method [4]. There are many available examples for one treelet, and also, there are many possible treelet combinations. The best combination is detected by log-linear decoding model with features described in Section 3.

In the example in Figure 1, four examples are used. They are combined and finally we can get the output dependency tree. We call the outside nodes of the actually used nodes as "bond" nodes. The bond nodes of one example are replaced by the other example, and thus two examples can be combined.

We also addresses two typical characteristics of patent documents which often reduce the translation quality. One is that the documents contain huge amount of technical terms and the other is that the sentences are very long on average. The special treatments for the patent translations are introduced in Section 4.

2. BAYESIAN TREELET ALIGNMENT MODEL BASED ON DEPENDENCY TREES

Alignment accuracy is crucial for providing high quality corpus-based machine translation systems because translation knowledge is acquired from an aligned training corpus. For similar language pairs, alignment accuracy is high. Less than 10% alignment error rate (AER) for French-English has been achieved by the conventional word alignment tool GIZA++, an implementation of the alignment models called the IBM models [1], with some heuristic symmetrization rules. However, for distant language pairs such as English-Japanese, the conventional alignment method is quite inadequate (achieving an AER of about 20%).

There are two main issues in a word alignment task for distant language pairs: one is the word order difference, while the other relates to function words. The word order issue has to some extent been solved by using word dependency trees in the alignment model [9]. Most of the remaining alignment errors are related to function words such as English articles and Japanese case markers [12] because they do not have counterparts in the other language. As an example, most

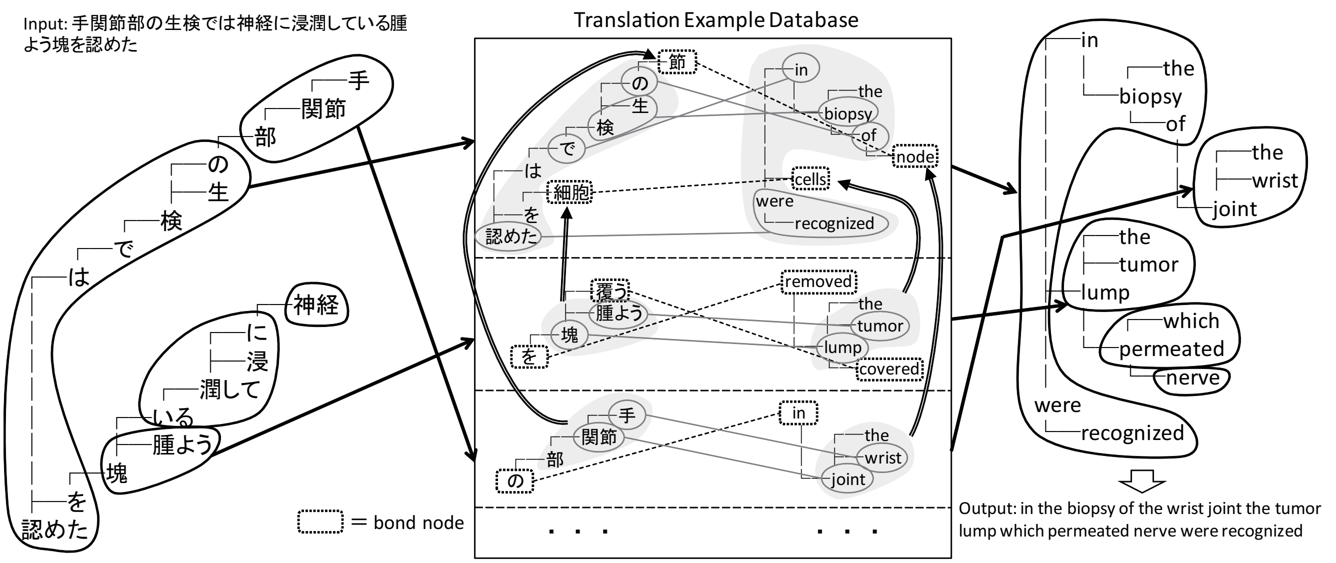


Figure 1: An example of Japanese-English translation.

of the errors in the top of Figure 2 are related to function words: "has" and "は (*topic-marker*)" in example (A), and "although", "は (*topic-marker*)", "を (*ACC*)" and "が (*but*)" in example (B).

To overcome the alignment errors related to function words, we exploit the treelet alignment model by bilingual generation and monolingual derivation based on dependency trees [11]. If there is a direct translation for a function word, these words should be aligned with each other. For function words that do not have any counterparts, the conventional model is supposed to align them to NULL, but it does not always work well. They are often aligned to some words incorrectly. In contrast with the conventional model, our model *derives* such function words from content words in their own language. The derivation probabilities used in our proposed model are estimated from a large monolingual corpus for each language. Thus, we do not require a large parallel corpus. With this derivation model, we can reduce alignment errors for function words, which leads to a better translation resources such as a phrase table, which is acquired from a word-aligned parallel corpus. The bottom of Figure 2 shows the alignment results by the treelet alignment model with monolingual derivation. The model reduced the alignment errors for unique function words by deriving them monolingually, and found correct alignments which the baseline system failed to find.

3. TREE-BASED TRANSLATION

As a tree-based translation method, we adopt example-based machine translation system [10]. In this section, we briefly introduce the translation procedure in the EBMT system.

3.1 Retrieval of Translation Examples

The input sentence is converted into the dependency structure as in the parallel sentence alignment. Then, for each treelet, available translation examples are retrieved from the example database. Here the word "available" means that all the words in the focusing input treelet appear in the source tree of the example, and the dependency relations between the words are same. We use the fast, on-line tree retrieval technique [4] to get all the available examples from huge training corpus.

3.2 Selection of Translation Examples

We find the best combination of examples by tree-based log-linear model with features shown below:

- **Size of examples**
- Translation probability
- Root node of examples
- Parent node
- Child nodes
- Bond nodes
- NULL-aligned words
- Language model

Among the features, an important one is "Size of examples". Translations with larger examples can achieve higher quality because translations inside the examples are stable.

3.3 Combination of Translation Examples

When combining examples, in most cases, *bond nodes* are available outside the examples, to which the adjoining example is attached. Using the bond information, we don't need to consider word or phrase orders. Bond information naturally resolve the reordering problem. Figure 1 is an example of combining translation examples. The combination process starts from the example used for the root node of the input tree (the first one in Figure 1). Then the example for the child node of the treelet covered by the initial example is combined (the second and third examples). When combining the second example to the first one, "細胞 ↔ cells" is

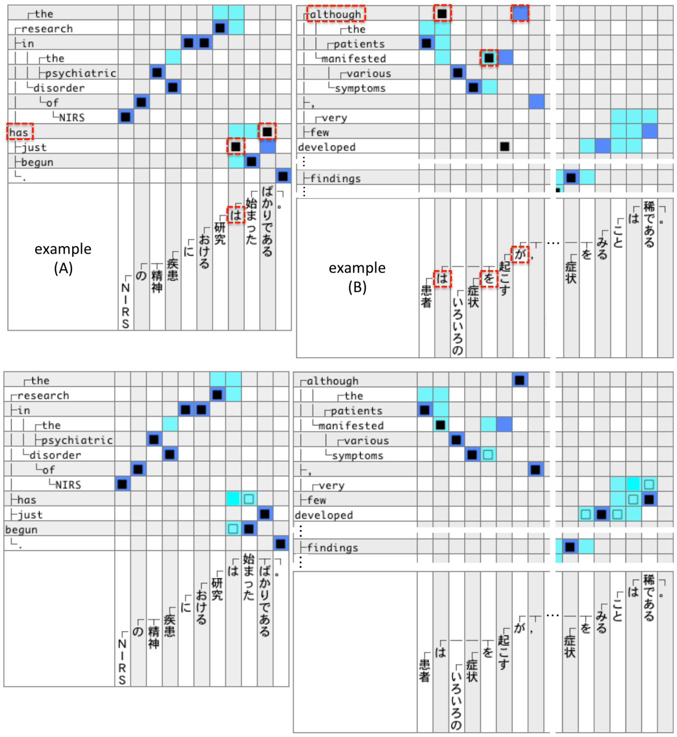


Figure 2: Alignment results of [9] (top) and [11] (bottom). The solid and open squares depict core alignments and derivations respectively, while dark blue (Sure) and light blue (Possible) cells denote gold-standard alignments . Cells demarcated by dotted lines are alignment errors related to function words.

used as bond node, and for the third example, "節 ↔ node" is used as bond node. The combination repeated until all the examples are combined into one target tree. Finally, output target sentence is generated from the tree structure.

Note that there are NULL-aligned nodes in the examples (the nodes which are not circled, such as ' は', ' を', ' 部 (*part*)' and articles in English).

4. SPECIAL TREATMENTS FOR PATENT TRANSLATION

Patent documents have typically two characteristics: one is that the documents contain huge amount of technical terms and the other is that the sentences are very long on average. These characteristics often derive word segmentation errors and parsing errors. To reduce these errors, we perform three types of special treatments.

4.1 Japanese Lexicon Acquisition

Japanese technical terms are rarely included in the dictionaries of the ordinary morphological analyzers. We extract such technical terms using the method of [8] from the Japanese sentences of the training corpus. Some examples of the acquired lexicon are listed in Table 1. The acquired lexicon includes 2 adjectives, 37 verbs and 830 nouns.

4.2 English Compound Noun Extraction

Table 1: Automatically extracted Japanese lexicon from patent sentences.

Lexicon	POS	form
急峻だ	adjective	ナ形容詞
不溶だ	adjective	ナ形容詞
嵌込む	verb	子音動詞マ行
組付ける	verb	母音動詞
圧送	verbal noun	N/A
移載	verbal noun	N/A
隔板	noun	N/A
支持桿	noun	N/A

English compound nouns often includes verbs inside, and the parsers sometimes incorrectly analyze such verb as the main verb of the sentence. Figure 3 shows an example of this case. The compound noun "the plate support member 23" includes verb "support" and the parser analyze it as the main verb. This type of parsing errors are critical to our EBMT system because our system highly depends on the parsing results, and the parsing errors easily lead to incorrect translations.

To reduce the parsing errors of English conpound nouns, we automatically extract compound nouns in English sentences using the alignment results and Japanese parsing results. Japanese compound nouns are relatively easy to detect and Japanese parsers correctly analyze the compond nouns. Therefore, English compond nouns can be acquired using the Japanese compound noun infomation and the alignment results.

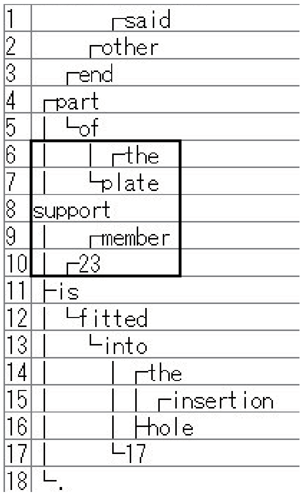
The extracted compound nouns are concatenated into a single word like "the-plate-support-member-23", then the sentences are again parsed. After the parsing, concatenated compound nouns are divided into pieces and word dependency subtrees are constructed as if words depend on the next word.

As for the English input sentences, we aggregate the compound nouns extracted from parallel training corpus and construct a list of compound nouns, and find the compound nouns in the input sentences by longest match strategy using the list. Note that low frequency compound nouns are discarded to keep the precision of detecting compound nouns.

4.3 PP-attachment Modification

English prepositional phrases often have ambiguities on choosing their parent phrase (known as PP-attachment ambiguities). These ambiguities cause both alignment and translation errors for tree-based translation methods such as our EBMT system. Fortunately, Japanese sentences have less ambiguities, furthermore, the Japanese dependency parser KNP [6] which we used for analyzing Japanese sentences can disambiguate the dependencies using case frame information. Therefore, using the Japanese-side dependency trees and the alignmet results, we modifies the parent of English prepositional phrases so as to make the dependency relations similar to those of Japanese sentences. We re-run the alignemnt module on the modified dependency trees.

For input English sentences, we have no idea to resolve the PP-attachment problem, thus we did not do anything. This is our future work.



```
1  │      ┌said
2  │    ┌other
3  │  ┌end
4  │┌part
5  ││└of
6  ││  │┌the
7  ││  └plate
8  │support
9  ││   ┌member
10 ││┌23
11 │├is
12 ││└fitted
13 ││  └into
14 ││    │┌the
15 ││    ││┌insertion
16 ││    │├hole
17 ││    └17
18 │└.
```

Figure 3: An example of the parsing error of English compound noun.

Table 2: NTCIR-10 intrinsic evaluation result

	J->E			E->J		
	RIBES	BLEU	Adeq.	RIBES	BLEU	Adeq.
Moses	69.72	28.56	2.81	72.31	32.98	2.69
KYOTO	67.24	24.01	2.74	72.52	26.85	2.50

5. NTCIR-10 PATENTMT RESULTS

We used the EBMT system described above for NTCIR-10 PatentMT [5]. English sentences were converted into phrase structures using Charniak's nlparser [2], and then they were transformed into dependency structures by rules defining head words for phrases [3]. Japanese sentences were converted into dependency structures using the morphological analyzer JUMAN [7] and the dependency analyzer KNP [6].

5.1 Official Results

Table 2 shows the formal run evaluation result of our KYOTO system compared to the Moses (BASELINE1) system. The automatic evaluation scores of Kyoto system, especially BLEU scores, are much inferior to Moses system, however, the Adequacy scores are competitive. As often mentioned, the automatic evaluation scores and the human evaluation scores are not necessarily correlated.

Table shows the chronological evaluation results. In the chronological evaluation, the same test set to the previous NTCIR-9 PatentMT is used, thus the participants can see the improvements of their systems from the last workshop. From the results, we can see very large improvement of the translation quality of our system. This is mainly because the improvement of word alignment accuracy explained in Section 2. However, we do currently not pay much attention to the decoder, and did not conduct the feature selection nor parameter tuning, and did not heavily use the language model. We believe our system becomes much better in near future after careful implementation of the decoder.

5.2 Effect of Special Treatments

Table 4 shows the effect of special treatments for patent documents, explained in Section 4, to the translation quality. The Japanese lexicon acquisition has good effect on both Ja-En and En-Ja translations. The English compound noun extraction has good effect on the En-Ja translation because this treatment reduces the parsing errors of input sentence. The English PP-attachment modification has almost no effect on the translations, even for the En-Ja translation. This is because the English input sentences are not modified by this treatment although those in training corpus are modified. Therefore, the inconsistency of PP-attachment between training and test sentences may occur. We need to consider a way of modifying the PP-attachment of input English sentences.

Table 3: NTCIR-10 chronological evaluation result

	J->E		E->J	
	RIBES	BLEU	RIBES	BLEU
Moess	70.07	28.47	72.44	32.10
NTCIR-9	65.15	21.49	66.11	24.59
NTCIR-10	69.18	24.65	72.35	26.52

Table 4: Effect of special treatments measured on the NTCIR-10 intrinsic test set

	J->E		E->J	
	RIBES	BLEU	RIBES	BLEU
baseline	69.60	23.41	72.77	25.47
+Ja lexicon	69.48	24.17	73.03	25.80
+En compound noun	69.53	23.89	73.39	25.80
+En PP modification	69.54	23.84	73.39	25.72

6. CONCLUSION

In this paper, we explained our linguistically-motivated translation framework which is composed of treelet alignment model by bilingual generation and monolingual derivation based on dependency tree structures, and example-based translation method where the examples are expressed in dependency tree structures.

Although our EBMT system basically can generate adequate and fluent translations, we could not achieve satisfactory results in the formal run because we did not pay much attention to the decoder. In near future, we will sophisticate the implementation of the decoder and see the improvement of translation quality.

7. REFERENCES

[1] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Association for Computational Linguistics*, 19(2):263–312, 1993.

[2] E. Charniak and M. Johnson. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180, 2005.

[3] M. Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, 1999.

[4] F. Cromieres and S. Kurohashi. Efficient retrieval of tree translation examples for syntax-based machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 508–518, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[5] I. Goto, K. P. Chow, B. Lu, E. Sumita, and B. K. Tsou. Overview of the patent machine translation task at the ntcir-10 workshop. In *Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-10)*, 2013.

[6] D. Kawahara and S. Kurohashi. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 176–183, New York City, USA, June 2006. Association for Computational Linguistics.

[7] S. Kurohashi, T. Nakamura, Y. Matsumoto, and M. Nagao. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of The International Workshop on Sharable Natural Language*, pages 22–28, 1994.

[8] Y. Murawaki and S. Kurohashi. Online acquisition of Japanese unknown morphemes using morphological constraints. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 429–437, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[9] T. Nakazawa and S. Kurohashi. Bayesian subtree alignment model based on dependency trees. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 794–802, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.

[10] T. Nakazawa and S. Kurohashi. Ebmt system of kyoto team in patentmt task at ntcir-9. In *In Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-9)*, pages 657–660, 2011.

[11] T. Nakazawa and S. Kurohashi. Alignment by bilingual generation and monolingual derivation. In *Proceedings of COLING 2012*, pages 1963–1978, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[12] X. Wu, T. Matsuzaki, and J. Tsujii. Effective use of function words for rule generalization in forest-based translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 22–31, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.