EBMT System of KYOTO Team in PatentMT Task at NTCIR-9

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 task at NTCIR-9. When translating very different language pairs such as Japanese-English and Chinese-English, it is very important to handle sentences in tree structures to overcome the difference. Some works incorporate tree structures in some parts of whole 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 Bayesian phrase alignment model on dependency trees and example-based translation.

Keywords

Syntactic EBMT, Tree-based, Bayesian Sub-tree Alignment

Team Name

KYOTO

Subtasks/Languages

Japanese to English, English to Japanese, Chinese to English

External Resources Used

GIZA++, JUMAN, KNP, nlparser, CNP

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 Bayesian subtree alignment model based on dependency trees. The details are briefly shown in Section 2. It is a kind of tree-based reordering model, and can capture non-local reorderings which sequential word-based models cannot often handle properly. In the translation, we adopt an example-based machine translation (EBMT) system [5] which is very conformable to the tree structures. EBMT can handle examples which are discontinuous as a word sequence, but continuous structurally. Accordingly, EBMT can quite naturally handle syntactic information. 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 Bayesian subtree alignment model. Note that both source and target sides of all the examples are stored in dependency tree structures. Using the example database, new input sentence is translated. The input sentence is also parsed and transformed into dependency structure. For all the arbitrary sub-trees, available examples are searched in the example database. This step is the most time consuming part, and we exploit a fast tree retrieval method[3]. Of course there are many available examples for one sub-tree, and also, there are many types of sub-tree combinations. We search the best combination 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. Using the bond information, we don’t need to consider word or phrase orders. Bond information naturally resolve the reordering problem.

2. BAYESIAN SUBTREE 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 distant language pairs such as English-Japanese or Chinese-Japanese, the word sequential models such as IBM models are quite inadequate (about 20 to 30 % AER), and therefore it is important to improve the alignment accuracy itself. The differences between languages can be seen in Figure 2, which shows an example of English-Japanese. The word or phrase order is quite different for these languages. Another important point is that there are often many-to-one or many-to-many correspondences. For example, the Japanese noun phrase “受 光 素子” is composed of three words, whereas the corresponding English phrase consists of only one word “photodetector”, and the English function word “for” corresponds to two Japanese function words “に は”. In addition, there are basically no counterparts for the English articles (a, an, the). Figure 3 shows the alignment results from bi-directional GIZA++ together with a combination heuristic called grow-diag-final-and for the same

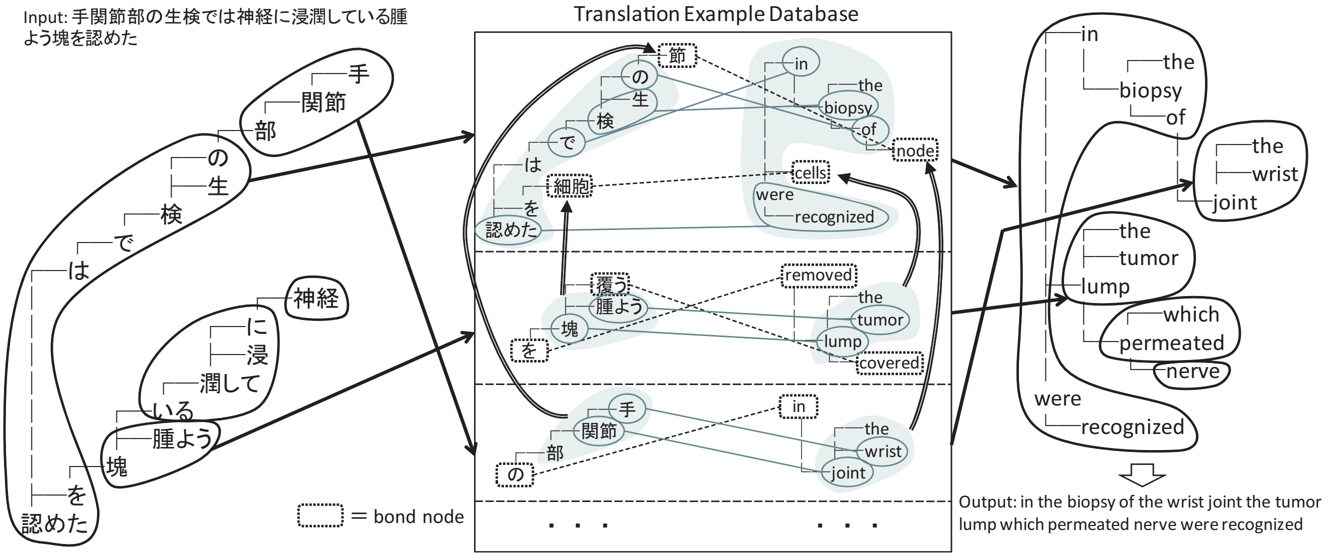

Figure 1: An example of Japanese-English translation.

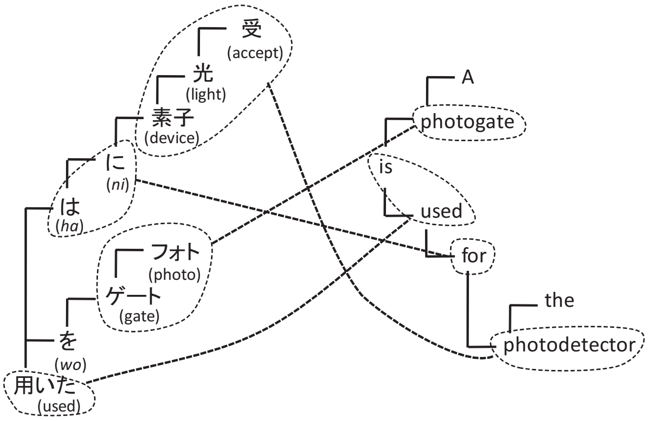

Figure 2: Example of dependency trees and alignment of subtrees. The root of the tree is placed at the extreme left and words are placed from top to bottom.

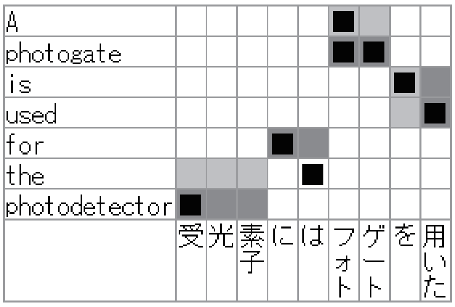

Figure 3: Alignment results from bi-directional GIZA++. Black boxes depict the system output, while dark (Sure) and light (Possible) gray cells denote gold-standard alignments.

sentence pair given in Figure 2. The system failed to align some words in the Japanese noun phrase, and incorrectly aligned "the ↔ は ". The word sequential model is prone to many such errors even for short simple sentences of a distant language pair.

Even if the word order differs greatly between languages, phrase dependencies tend to hold between languages. This is also true in Figure 2. Therefore, incorporating dependency analysis into the alignment model is useful for distant language pairs. We exploit Bayesian subtree alignment model based on dependency trees [6]. This model incorporates dependency relations of words into the alignment model and define the reorderings on the word dependency trees. Figure 2 shows an example of the dependency trees for Japanese and English.

3. TREE-BASED TRANSLATION

As a tree-based translation method, we adopt example-based machine translation system [5]. 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 sub-tree, available translation examples are retrieved from the example database. Here the word "available" means that all the words in the focusing input sub-tree 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 [3] 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". We think 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. 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 sub-tree 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 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. NTCIR-9 PATENTMT TASK

We used the EBMT system described above for NTCIR-9 Patent Translation Task. The detail of the task is described in [4]. Table 1 shows the formal run evaluation result of our KYOTO system compared to the "BASELINE1" system. Unfortunately, the results includes some bugs in the example retrieve module and also in some other modules. We fixed them after the formal run and tested again in Japanese-to-English directions with NTCIR-9 formal run set. The results are shown in Table 1 with parenthesis.

Major translation errors of patent translation come from incorrect parsing results of technical terms in English sentences. In the left of Figure 4 shows an example of incorrect parsing result of an English sentence with a technical term. In the example, "the plate support member 23" is the technical term and "is fitted" should be the main verb, however "support" is analyzed as the main verb. Since our EBMT system highly depends on the parsing results, such parsing errors easily lead to translation errors. This problem can be solved by using monolingual technical term dictionary which is automatically acquired from context documents [7].

Some technical terms are also problematic for Japanese sentences. In patent documents, there are so many technical terms which end with numerals (right side of Figure 4[1]), and they are sometimes incorrectly parsed. In the example, "研磨テープ２１ (*polishing tape 21*)" should depend on "吹き付ける (*blow*)", but is analyzed as depending on "設置 (*set*)". This is because the numeral is regarded as the head of the noun phrase, and linguistic knowledge between the noun phrase and a verb is unavailable. This problem can be solved by regarding the closest child ("テープ (*tape*)" in the example) of the numeral as the head of the noun phrase.

Table 1: NTCIR-9 Official Evaluation Result (The BLEU score with parenthesis is the re-evaluation result after formal run.)

	J->E		E->J		C->E	
	BLEU	Adeq.	BLEU	Adeq.	BLEU	Adeq.
KYOTO	21.14 (22.89)	2.38	24.52	2.05	17.8	2.41
BASELINE1	28.95	2.61	31.58	2.60	30.7	3.29

As for Chinese-to-English translation, the BLEU score was much lower than than that of Japanese-to-English or English-to-Japanese. This is because, the parsing accuracy of Chinese sentence is much lower than Japanese and English, so currently it is hard for our EBMT system to achieve high quality translation in Chinese-to-English direction. Both the English and Japanese parsers used in the experiments can analyze sentences with over 90% accuracy, whereas the accuracy of the Chinese parser is less than 80% despite it being state-of-the-art in the world [2]. The parsing accuracy reported in this paper was obtained from an experiment using gold-standard word segmentation and POS-tags. Starting with raw sentences results in about 77.4% accuracy. This information was obtained from communication with the authors. However, the Chinese parsing must be improved in the long run, and also the translation quality of our EBMT system should be improved. One possible short-term solution for the parsing problem is to use the n-best parsing results in the model. Another kind of solution was proposed by Burkett et al. [1], who described a joint parsing and alignment model that can exchange useful information between the parser and aligner.

5. CONCLUSION

In this paper, we have proposed a linguistically-motivated translation framework which is composed of Bayesian subtree alignment model 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 failed to do careful treatment specialized for patent documents such as technical terms. In the future, we will further investigate the patent documents and find the way to better translation quality.

6. REFERENCES

[1] D. Burkett, J. Blitzer, and D. Klein. Joint parsing and alignment with weakly synchronized grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages

[1] English translation: *A blower 56 to blow cold air to the polishing tape 21 is set between the guide roller 54 and a semiconductor wafer 10.*

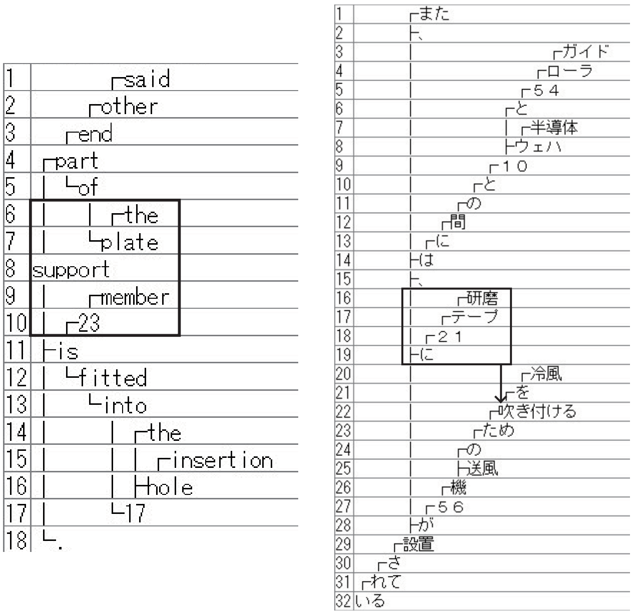

Figure 4: Parsing Error of English Technical Term (left) and Japanese Noun Phrase which Ends with Numeral (right).

127–135, Los Angeles, California, June 2010. Association for Computational Linguistics.

[2] W. Chen, D. Kawahara, K. Uchimoto, Y. Zhang, and H. Isahara. Dependency parsing with short dependency relations in unlabeled data. In *In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*, pages 88–94, 2008.

[3] 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.

[4] I. Goto, B. Lu, K. P. Chow, E. Sumita, and B. K. Tsou. Overview of the patent machine translation task at the ntcir-9 workshop. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, 2011.

[5] T. Nakazawa and S. Kurohashi. Fully syntactic ebmt system of kyoto team in ntcir-8. In *In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-8)*, pages 403–410, 2010.

[6] 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.

[7] T. Onishi, M. Utiyama, and E. Sumita. Reordering constraint based on document-level context. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 434–438, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.