Parsing Myanmar (Burmese) by Using Japanese as a Pivot

Chenchen Ding, Ye Kyaw Thu, Masao Utiyama, Eiichiro Sumita
Multilingual Translation Laboratory, NICT
3-5 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0289, Japan
{chenchen.ding, yekyawthu, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

As Myanmar (Burmese) and Japanese share similar syntactic structures, we explore an approach for parsing Myanmar by using Japanese as a pivot. Specifically, we first apply a statistical machines translation (SMT) system to translate Myanmar sentences into Japanese, and then we parse the Japanese translations. Finally, the Japanese syntactic structures are mapped to the original Myanmar sentences. As a state-of-the-art SMT system performs well on translating syntactically-similar languages and the parsing techniques for Japanese have been well-studied, we have satisfactory results in experiments.

1. Introduction

Syntactic parsing is one of the basic tasks in natural language processing (NLP), focusing on revealing the syntactic structures of sentences belong to particular natural languages. As the structures of natural languages have much more ambiguities than artificial programing languages, the parser of natural languages cannot be easily realized by simple rules and deterministic algorithms. Instead, annotated corpora (i.e., tree-banks) and data-driven statistical approaches are studied and applied for academic researches and practical applications. As the construction of annotated corpus costs much in both time and money, it turns to be a long-term project, especially for those understudied, low-resource languages, such as Myanmar.

In this paper we explore an approach for parsing Myanmar without annotated Myanmar corpus, but use Japanese, a well-studied language, as a pivot. The motivation of this work based on three facts: 1) Japanese and Myanmar have quite similar syntactic structures; 2) we have considerable Japanese-Myanmar parallel data and the state-of-the-art statistical machine translation (SMT) techniques can give satisfactory translation performance between the two languages; 3) the available Japanese morphological analysis tool and parser can give satisfactory results. According to the above-mentioned points, the specific process of parsing Myanmar sentence in this paper is composed of three steps: 1) to translate a Myanmar sentence in to Japanese sentence using a trained SMT system; 2) to parse the Japanese sentence; 3) to map the structure of the Japanese sentence into the original Myanmar sentence. For step 1) and 2), off-the-shelf SMT system and Japanese processing tools are used. For step 3), we propose a simple approach to establish the correspondence structures between Japanese and Myanmar.

In order to examine the performance of the proposed approach, we conducted experiments on the *basic travel expressing corpus* (BTEC) with manual checks on samples from a test set. We mainly focus on evaluation of two aspects: 1) the correctness of the parsing on the Japanese translation, and 2) the correctness of the mapped struc-

ture on the original Myanmar sentence. The experimental results illustrate, that we can actually obtain relatively correct syntactic structures on Myanmar sentences by using Japanese as a pivot.

2. Proposed Approach

As mentioned, the process of translating Myanmar sentences into Japanese and parsing Japanese are only using off-the-shelf tools, we mainly describe the mapping process of establishing the final Myanmar sentence structure in this section. The specific settings and details of SMT and Japanese processing will be presented in the following section of experiment.

Because Japanese and Myanmar are typical head-final languages with abundant case-markers and particles to illustrate the relations between phrases within a sentence, their word orders are relatively free as long as the head-final restriction is satisfied. Consequently, dependency relation on chunk-level is a more suitable grammatical formulation than constituency-based analysis on this kind of language. In this work, the Japanese sentences are first parsed to obtain their chunk dependencies, and then the chunks and dependency between chunks are mapped to the Myanmar-side successively. We describe the two mapping steps in the following sub-sections respectively.

2.1. Chunk Mapping

Fig. 1 illustrates the process of chunk mapping between Japanese and Myanmar. With the help of token alignment between input and output generated by the SMT system, we can get a rough correspondence of chunks. Specifically, each chunk on Japanese-side is examined to figure out the corresponding cover range on the Myanmar-side. Then we use two heuristic rules to refine the rough correspondence in order to get a clean chunk segmentation on the Myanmar-side. They are: 1) to merge overlapped chunks, and 2) to appending left-out words to it previous chunk. Rule 1) is a conservative process to handle the disagreement of chunk segmentation and translation on Japanese, which may be caused by alignment errors in the SMT system or by inherent difference in expression between the two languages. We only prefer under-segmenting rather than over-segmenting for the Myanmar sentences. We further use rule 2) to group unaligned words because they are usually functional particles positioned after the content word they are related to. Generally, the chunk mapping process from Japanese to Myanmar will reduce the number of chunks because we want to ensure the Myanmar sentence not to be over-segmented.

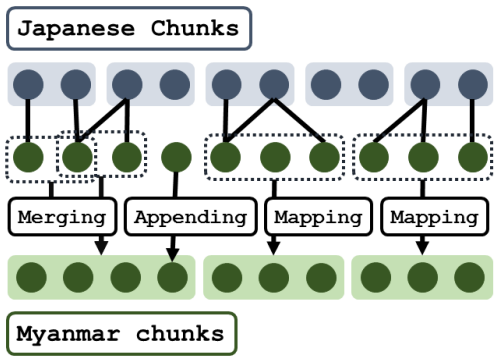

Figure 1. Process of chunk mapping. Dots stand for tokens and solid boxes stand for chunks. The first row stands for the Japanese sentences; the second and third rows stand for original Myanmar sentence. Between the first and second rows, corresponding tokens in translation are linked.

2.1. Dependency Mapping

Fig. 2 illustrates the mapping of dependency relations after the mapping of chunks. Basically, the head-modifier chunk pairs on Japanese-side will be mapped to Myanmar-side directly. As

mentioned, the number of chunks may be reduced on the Myanmar-side. So, for 1) merged chunks, only the dependency arc of the coresponding right-most Japanese chunk will be retained and other related arcs are deleted; for 2) disappeared chunks on Japanese-side, all the dependency arcs rooted on them will be passed to their first undisappeared ancester chunk, and all the arcs pointed to them will be deleted. After the above two modifications, we can obtain a detemined mapping of the dependency arcs.

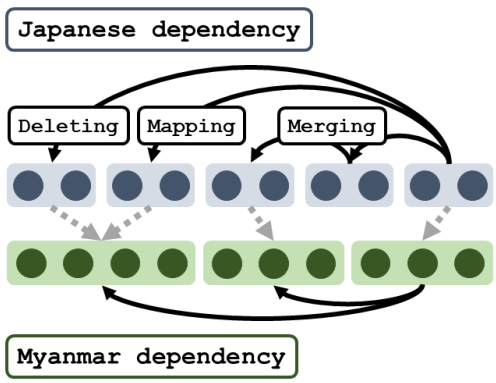

Figure 2. Process of dependency mapping. The meaning of dots and boxes are identical to Fig. 1. The relation between chunks of the two sides is marked by gray arrows. The dependency relation of chunks are noted by arcs.

3. Experiment

3.1. Data and Tools

We investigated the performance of our proposed approach using the *basic travel expressing corpus* (BTEC) [1]. The corpus is composed of three sets of training, development, and test, containing 457,249 sentences, 5,000 sentences and 3,000 sentences respectively. The training and develop sets were used to train a standard phrase-based (PB) SMT system to translate Myanmar into Japanese. The test set was used for testing and evaluating the Myanmar parsing approach proposed in this paper.

For the PB SMT, we used MOSES[1] [2] with default settings in training and decoding. The language model used in SMT was an interpolated modified Kneser-Ney discounting 5-gram model, trained on the Japanese part of the training set by SRILM[2] [3]. The Myanmar sentences were segmented into words using an in-house CRF-based tool for the SMT system. For the output Japanese sentences from the SMT system, we used MeCab[3] with IPA dictionary for segmenting and CaboCha[4] [4] for chunking and parsing.

3.2. Evaluation

First, we report the performance of the PB SMT system on Myanmar-to-Japanese translation used in experiments. We tested different setting and the results are listed in Table 1.

Table 1. Performance of PB SMT on Myanmar-to-Japanese translation.

DL	Lex.-Reo.	BLEU	RIBES
0	no	**38.5**	.812
0	yes	38.2	.812
3	yes	38.2	.812
6	yes	**38.5**	**.814**

The DL in Table 1 means the distortion-limit used in decoding, and the column of Lex.-Reo. shows whether the lexicalized orientation reordering is used. Two measures, the BLEU score [5] and RIBES [6] are reported in Table 1. We observed a purely monotone translation (DL=0)

[1] http://www.statmt.org/moses/

[2] http://www.speech.sri.com/projects/srilm/

[3] http://taku910.github.io/mecab/

[4] http://taku910.github.io/cabocha/

without reordering model can give a good results and a larger DL could hardly improve the performance, although using a DL of 6 may give a slightly higher RIBES. The result is reasonable due to the similarity of the two languages. So, we simply used the settings of DL=0 and no reordering model for the PB SMT system in experiments.

We evaluated the accuracy of chunking and dependency on both Japanese and Myanmar. Because the Japanese sentences are output by the PB SMT system, sometimes they may become unnatural, which will cause the parsing error and finally affect the dependency mapping to Myanmar.

For the evaluation on chunks, we only concentrated on the part within the chunks to judge whether they are meaningful. Hence, Japanese chunks with unknown words are all labeled "not correct". On the Myanmar side, the chunks are generally long phrases, which may cover different clauses caused by the merging step in our approach. This kind of Myanmar chunks are also labeled as not correct. For the evaluation on dependency, the head-modifier pair of chunks are examined to see whether they have syntactic relations. If a chunk is meaningless, either in the translated Japanese and original Myanmar, the dependency around it may also be wrong. However, some dependency results are correct on wrong chunks, because of the good performance of the parser.

Table 2. Chunk and dependency accuracy on Japanese and Myanmar sentences.

	Japanese	Myanmar
chunk	91.6% (97.3%)	95.7%
dependency	98.6%	95.8%

The percentage of chunk and dependency accuracy on 500 randomly sampled Japanese / Myanmar sentence pairs from the test set are listed in Table 2. The percentage in brackets of Japanese-chunk does not take those chunk with unknown words as "not correct". As to the denominator of the percentage, i.e., the number of chunks, the original 500 Japanese sentences contain 1,604 chunks and after mapping to the Myanmar side, it reduced to 1,344 chunks.

3.3. Discussion

From the evaluation results in Table 2, it can be observed that the final chunk and dependency accuracy is over 95%. The high numerical results demonstrate the proposed approach actually generates reasonable paring results on Myanmar sentences. On the translated Japanese sentences, the unknown Myanmar words did not affect the approach much, because the Japanese parsing more relies on those functional morphemes, which are usually translated well. A translation and mapping example is given in Fig. 3.

5. Related and Future Work

To solve the parsing problem of low-resource languages by sharing features with high-resource language has been studied in recent research [7, 8]. However, the "low-resource languages" referenced in these studies are actually not really "low-resource" but only having less resource than those English large data set. For example, in the work of [9], the Irish language was taken as a case-study and referred to as a low-resource language; however, there are well-annotated Irish data which were used in this study. Hence, the Myanmar language is more proper to be referred to as "no-resource" because presently we still have no well-developed tree-bank for it, or even no refined part-of-speech tag set to describe Myanmar.

Recently, a line of research attempts to model different languages using a universal syntactic formulation [10, 11]. We think this will be a proper framework for annotated corpus building and sophisticated Myanmar-oriented processing, on which we are working presently.

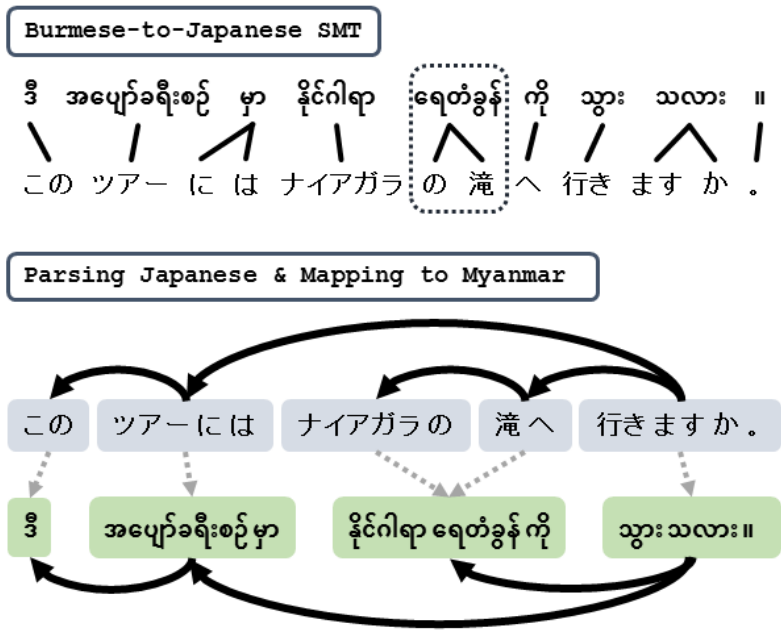

Figure 3. Example of translation and mapping in our experiment. The part in dashed box on Myanmar-to-Japanese SMT is a noise brought by the SMT system, which causes two Japanese chunks merged into one Myanmar chunk in the mapping step. Even though, the mapped dependency relation on Myanmar is correct because the Japanese genitive case-marker "の" have no Myanmar corresponding translation in this example.

5. Conclusion

In this paper, we proposed an approach to parse Myanmar sentences by using Japanese as a pivot. The chunk and dependency structures of a Myanmar sentence are mapped from its Japanese translation with the help of a PB SMT system and a Japanese parser. Experimental results on a basic travel expression corpus were examined manually and illustrated the proposed approach performed satisfactory. We hope the work reported in this paper can play an auxiliary role in the future Myanmar tree-bank construction.

Acknowledgement

We thank Dr. Win Pa Pa from NLP Lab., University of Computer Studies Yangon (UCSY) for her help in evaluating the experiment results.

References

[1] G. Kikui, "Creating Corpora for Speech-to-Speech Translation", In Proc. of INTERSPEECH2003, pp. 381-384.

[2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation". In Proc. of ACL2007, pp. 177-180.

[3] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit", In Proc. of ICSLP2002, pp. 901-904.

[4] T. Kudo and Y. Matsumoto, “Japanese Dependency Analysis Using Cascaded Chunking”, In Proc. of CoNLL2002, pp. 63-69.
[5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A Method for Automatic Evaluation of Machine Translation”, In Proc. of ACL2002, pp. 311-318.
[6] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada, “Automatic evaluation of translation quality for distant language pairs”, In Proc. of EMNLP2010, pp. 944-952.
[7] L. Duong, T. Cohn, S, Bird and P. Cook, “Low Resource Dependency Parsing: Cross-lingual Parameter Sharing in a Neural Network Parser”, In Proc. of ACL2015, pp. 845-850.
[8] L. Duong, T. Cohn, S, Bird and P. Cook, “A Neural Network Model for Low-Resource Universal Dependency Parsing”, In Proc. of EMNLP2015, pp. 339-348.
[9] T. Lynn, J. Foster, M. Dras, and L. Tounsi, “Cross-lingual Transfer Parsing for Low-Resourced Languages: An Irish Case Study”, In Proc. of the First Celtic Language Technology Workshop, pp. 41-49.
[10] J. Tiedeman, “Cross-Lingual Dependency Parsing with Universal Dependencies and Predicted PoS Labels”, In Proc of Depling 2015, pp. 340-349.
[11] J. Tiedeman, Ž. Agić, and J. Nivre, “Treebank Translation for Cross-Lingual Parser Induction”, In Proc. of CoNLL2014, pp. 130-140.