Phrase Reordering for Statistical Machine Translation Based on Predicate-Argument Structure

Mamoru Komachi, Yuji Matsumoto

Nara Insitute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi
Nara, 630-0192 Japan
{mamoru-k,matsu}@is.naist.jp

Masaaki Nagata

NTT Communication Science Laboratories
2-4 Hikari-dai, Seika-cho, Soraku-gun
Kyoto, 619-0237 Japan
nagata.masaaki@lab.ntt.co.jp

Abstract

In this paper, we describe a novel phrase reordering model based on predicate-argument structure. Our phrase reordering method utilizes a general predicate-argument structure analyzer to reorder source language chunks based on predicate-argument structure. We explicitly model long-distance phrase alignments by reordering arguments and predicates. The reordering approach is applied as a pre-processing step in training phase of a phrase-based statistical MT system. We report experimental results in the evaluation campaign of IWSLT 2006.

1. Introduction

Recently, phrase-based statistical machine translation model has become the mainstream in the machine translation community. Phrase-based approaches are capable of constructing better context-dependent word selection model than word-based approaches. Though the unit of translation is still under active development [1], there is no approach more widely used than phrase-based one.

Statistical machine translation, however, uses less linguistic knowledge such as syntax and semantics than conventional rule-based machine translation systems. For instance, the chunk-based approach in [2] does not rely on monolingual chunker and the hierarchical phrase-based approach in [1] does not use any kind of syntactic information except for a synchronous context-free grammar. Some SMT systems, however, try to incorporate syntactic knowledge, such as [3], yet it is hard to use it effectively as described in [4]. Another issue in statistical machine translation is reordering. Global reordering is essential to translation of languages with different word orders [5], and some aspects of global reordering in translation between German and English was stated in [6] and [7]. They used some heuristics to pre-process German corpus and reported successful results.

In this paper, we present a novel phrase reordering model based on a predicate-argument structure analyzer. Given predicate-argument structure information from the analyzer, source sentence is reordered according to match that of the target language. The translation model trained on a reordered corpus has more monotonic phrase alignments and gets longer phrase alignments. In order to cope with data sparseness problem, we combined the original corpus and the reordered corpus. We used the Pharaoh [8] beam search decoder and observed improvements in the BLEU and NIST scores, and the reordering model could be combined with any kind of phrase-based statistical machine translation method.

In the following sections, we first explain our translation model and phrase reordering model. We then report the experiments' results using our phrase reordering model based on predicate-argument structure.

2. Baseline Translation Model

We followed the noisy channel approach to machine translation. In this approach, we search for the target (English) sentence by maximizing the probability of the target sentence $\hat{e}$ given the source (foreign) sentence $\hat{f}$. By applying Bayes' rule, we can formulate the process as maximizing the product of $P(e)$ and $P(f|e)$.

$$\hat{e} = \operatorname*{argmax}_{e} P(e|f) = \operatorname*{argmax}_{e} P(e)P(f|e)$$

This equation shows that the source language is transformed into target language through a noisy channel, and the translation process is to decode the source sentence from the target sentence. Here, we call the prior probability $P(e)$ the language model, and the conditional probability $P(f|e)$ the translation model.

In the phrase-based translation model, the source sentence f is segmented into a sequence of I phrases, $\overline{f}_1^I$. Each source phrase $\overline{f}_i$ in $\overline{f}_1^I$ is translated into a a target phrase $\overline{e}_i$. The target phrases may be reordered. Phrase translation is then modeled by a probability distribution $\phi(\overline{f}_i|\overline{e}_i)$ and reordering of target phrases is modeled by a relative distortion probability distribution $d(a_i - b_{i-1})$, where a_i denotes the starting position of the source phrase phrase that was translated into the ith target phrase and b_{i-1} denotes the end position of the source phrase translated into the $(i-1)$th target phrase.

$$P(\overline{f}_1^I|\overline{e}_1^I) = \Pi_{i=1}^I \phi(\overline{f}_i|\overline{e}_i) d(a_i - b_{i-1})$$

Translation probability is obtained from the relative frequency of the source phrase given the target phrase aligned by the GIZA++ toolkit [9].

$$\phi(\overline{f}|\overline{e}) = \frac{\mathrm{count}(\overline{f}, \overline{e})}{\sum_{\overline{f}} \mathrm{count}(\overline{f}, \overline{e})}$$

where $\mathrm{count}(\overline{f}, \overline{e})$ gives the source phrase $\overline{f}$ aligned to the target phrase $\overline{e}$ in the parallel corpus.

The distortion model can be defined as follows with an appropriate value for the parameter α:

$$d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1} - 1|}$$

3. Phrase Reordering Model

Although the baseline phrase beam search decoder has succeeded to translate similar language pair such as English and French, it does not handle language pairs with significant syntactic differences such as Japanese and English. The problem here is that one syntactic and semantic unit in the source language might appear in a different position in the target language. Also, the fact that the word order is fairly free in Japanese makes it hard to align the arguments of predicates consistently.

Given that, we consider that it is important to model long distance phrase distortion correctly, and we try to solve this problem by using predicate-argument structures on the Japanese side.

3.1. Predicate-Argument Structure

In recent years, POS tagging and dependency parsing have achieved great accuracy. Predicate-argument structure analysis is the next step from syntax towards semantics.

The general problem of understanding text involves identifying semantic relations, "when" and "where" "who" did "what" to "whom".

Given a parsed sentence, the task of predicate-argument structure analysis is:

1. To identify predicates in the sentence, and
2. To determine their predicate-argument structures, and
3. To assign semantic roles to their arguments.

Predicates could be one of verbs, verbal nouns or adjectives.

In order to build a Japanese predicate-argument structure analyzer, we have been manually annotated Kyoto Text Corpus (Version 3.0) [10] with three case relations [11], GA, WO and NI, which roughly correspond to nominative, accusative and locative cases, respectively.[1] We then trained a predicate-argument structure analyzer [12] on our corpus, which assigns these three cases to the arguments of predicates given a sentence.

Figure 1 describes Japanese predicate-argument structure analysis of the following sentence:

住宗/address あ/WO-ACC ここ/here あ/NI-LOC 宗/write あ/PARTICLE 下さい/please

In this case, "宗/write あ/PARTICLE 下さい/please" is identified as a predicate, "住宗/address あ/WO-ACC" is assigned WO case, and "ここ/here あ/NI-LOC" is assigned NI case, respectively.

Our predicate-argument structure analyzer does not only use dependency information and explicit case markers, but also uses other features such as selectional preferences from several thesauruses and co-occurrences from newspapers.

It is important to note that the predicate-argument structure analyzer used here is a general and multipurpose analyzer that could deal with grammatical relations.

In Japanese, there is a topic marker WA, which introduces a topic in the sentence and overrides an explicit case marker GA and WO. [13] reports that WA accounts for 13.2% of all postpositions in Kyoto Text Corpus, and thus we can improve reordering model by recovering implicit case relations. In contrast to clause reordering based on syntactic information, we can disambiguate an argument with a topic marker or recover implicit case marker in spontaneous speech by using predicate-argument structure.

Although our predicate-argument structure analyzer is not capable of detecting exophora, or omitted predicate arguments outside the document, it can identify zero-anaphora, or omitted predicate arguments inside the document. Resolving exophora needs domain knowledge to some extent, but we do not have any tagged corpus on the domain concerning the evaluation campaign.

Our predicate-argument structure achieves a precision of 80.5% and a recall of 82.5% for the assignment of WO argument on predicates of verbal nouns [11].

3.2. Phrase Reordering

We uses the output of predicate-argument structure analyzer to reorder the arguments in Japanese sentences:

1. Find predicates (verbs, adjectives and verbal nouns).
2. For each predicate, reorder its arguments as follows:
    - Move nominative case (GA) chunk and its dependent trees to just before the predicate.
    - Move accusative case (WO) chunk and its dependent trees to right after the predicate.
    - Move locative case (NI) chunk and its dependent trees after the predicate.

[1] http://cl.naist.jp/nldata/corpus/

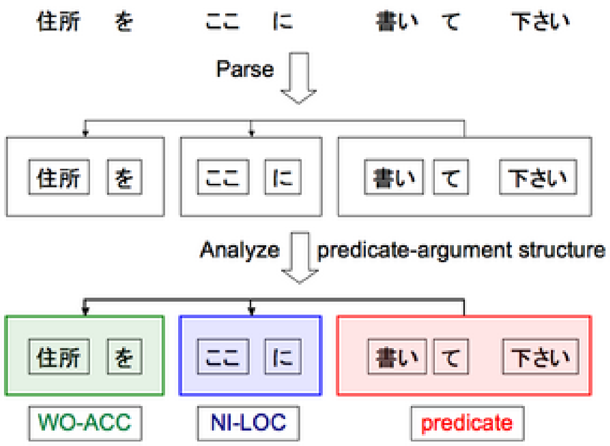

Figure 1: An example of Japanese predicate-argument structure analysis

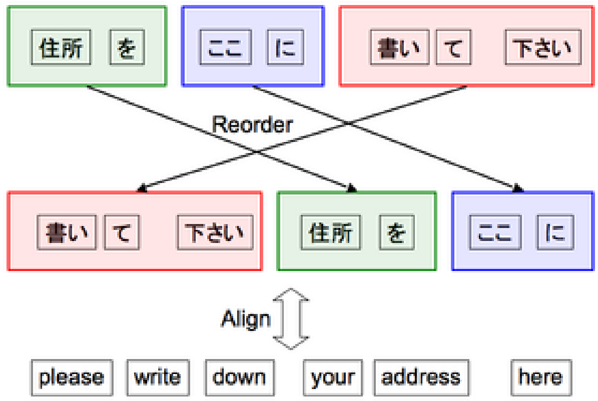

Figure 2: An example of phrase reorder

If there are multiple predicates which share an argument, only the first predicate is concerned with this movement.

Figure 2 illustrates the phrase reordering. “住宗/address あ/WO-ACC” and “ここ/here あ/NI-LOC” is moved after the predicate “宗/write あ/PARTICLE 下さい/please, ” and then the reordered sentence is aligned to the corresponding English sentence, “please write down our address here.” After reordering, there remains only one local crossing alignment (between “宗/write” to “write” and “下さい/please” to “please”).

These reordering steps are similar to the clause restructuring process described in [6]. The difference between their method and ours is that they used several heuristics to improve alignments of verbs while we are concentrating on aligning arguments. The same remark also applies to [7], and both of them use syntactic information for reordering model while we use semantic information in addition to syntactic information.

The reordering model presented above is not specific to Japanese and is applicable to other source languages which have relatively free word order and topicalization such as German, Russian, and other East Asian languages. However, most of the semantic role labeling systems use machine learning techniques which require semantically annotated corpus such as PropBank [14] and Chinese PropBank [15], and we also need a syntactic parser to start developing predicate-argument structure analyzer on top of it.

4. Experiments and Discussions

4.1. Corpus and Tools

We participated in Open Data Track in Japanese-English translation because we have built only Japanese predicate-argument structure analyzer and thus source language is limited to Japanese in our phrase reordering model.

We used ChaSen [16] for Word segmentation and POS tagging for Japanese. We did not use the original word segmentation information of Japanese because we used another POS tagger, ChaSen, instead. Dependency parsing was done by CaboCha [17]. We used tokenizer.sed from LDC to tokenize English sentences, and MXPOST [18] for POS tagging. Word translation probabilities were calculated by GIZA++ [9]. English words were lowercased for training and testing. We used a back-off word trigram model for the language model. It is trained on the lowercased English side of the parallel corpus by Palmkit [19].

We first manually aligned English and Japanese sentences and obtained parallel corpus of 45,909 Japanese-English sentences from 39,953 conversations. We then reordered Japanese sentences by using the predicate-argument structure analyzer. We added phrase reordered sentences to training set because phrase reordered corpus alone degrades translation results partly because our predicate-argument structure analyzer is trained only on news wire sources, and thus fails to identify the arguments of predicates correctly on text from a travel domain corpus.

The number of reordered sentences was 18,539 out of 45,909 (40.4%). The number of sentences which has crosssing alignment in the training corpus is 40,290, and after building combined corpus and running GIZA++ on the corpus, the number of sentences which has crossing alignments becomes 39,979. With regard to crossing alignments for each sentence, 33,874 sentences has less crossing alignments after reorder while 7,959 sentences gets more crossing alignments partly due to mis-classification of predicate-argument structure. These figures are shown in Table 1.

	number of sentences
reordered sentence	18,539
improve alignment	33,874
degrade alignment	7,959
crossing alignment	39,979
total	45,909

Table 1: Statistics on training corpus

4.2. Results

We compared our phrase reordering model with the state-of-the-art phrase translation method. The baseline system is based on WMT 2006 shared task baseline system [2]. We manually aligned English and Japanese corpus. Word segmentation and Japanese tokenization were done by the organizer, and true casing was performed by the scripts provided by the organizer. We used the same back-off word trigram model trained for proposed system for the language model. We used GIZA++ for alignment, created translation model by Pharaoh with default parameter configuration including phrase translation probability, lexical translation probability, phrase penalty, and phrase distortion probability.

For testing, we allowed pharaoh to reorder words during decoding, because we were interested in the question whether this system can be improved by explicit modeling of word order differences.

Table 2 shows the NIST and BLEU scores for development set 3, ASR 1-best recognition result and correct transcription of evaluation campaign in Japanese-English translation. If we reorder development set 3 for testing, the translation accuracy gets worse (see "reorder devset3" column in Table 2). This is because the predicate-argument structure analyser does not consistently identify arguments, and pharaoh handles reordering better than our reordering model during decoding. All the other testsets in the table were not reordered.

We used a minimum error rate training (MERT) tool provided by CMU [20] with 500 normal order sentences to tune parameter weights for the Pharaoh decoder after training both baseline system and proposed system.

However, it turned out that it does not always improve the translation accuracy in our model. The features we used were:

- Phrase translation probability
- Lexical translation probability
- Phrase penalty
- Phrase distortion probability

This is probably because it is difficult to optimize these features from small set of corpus, but we could not get results on large set of corpus containing both normal ordered and reordered sentences due to the computational difficulty. It is worth studying how many reordered sentences we need to improve translation accuracy.

In the experiments, the BLEU and NIST scores for all systems that use the reordered corpus outperform the conventional translation method. We did not try any other phrase-based decoder on the reordered corpus. it would be interesting to apply other decoders to the corpus.

[2]http://www.statmt.org/wmt06/shared-task/baseline.html

5. Conclusion

In this paper, we presented a novel phrase reordering model based on a general predicate-argument structure analyzer. We observed that the phrase reordering model improves baseline phrase translation model. Although our predicate-argument structure analyzer does not achieve satisfactory accuracy, we could improve phrase alignments between syntactically different language pairs. There is still room for improvements on how to incorporate syntactic and semantic information effectively into a statistical machine translation framework.

6. References

[1] D. Chiang, "A Hierarchical Phrase-Based Model for Statistical Machine Translation," in *Proceedings of the 43rd Annual Meeting of the ACL*, 2005, pp. 263–270.

[2] T. Watanabe, E. Sumita, and H. Okuno, "Chunk-based Statistical Translation," in *Proceedings of the 41st Annual Meeting of the ACL*, 2003, pp. 303–310.

[3] K. Yamada and K. Knight, "A Syntax-based Statistical Translation Model," in *Proceedings of the 39th Annual Meeting of the ACL*, 2001, pp. 523–530.

[4] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-Based Translation," in *Proceedings of the HLT-NAACL*, 2003, pp. 48–54.

[5] K. Ohashi, K. Yamamoto, K. Saito, and M. Nagata, "NUT-NTT Statistical Machine Translation System for IWSLT 2005," in *Proceedings of IWSLT*, 2005, pp. 128–133.

[6] M. Collins, P. Koehn, and I. Kucerova, "Clause Restructuring for Statistical Machine Translation," in *Proceedings of the 43rd Annual Meeting of the ACL*, 2005, pp. 531–540.

[7] S. Niessen and H. Ney, "Morpho-Syntactic Analysis for Reordering in Statistical Machine Translation," in *Proceedings of MT-Summit VIII*, 2001, pp. 247–252.

[8] P. Koehn, *Pharaoh: a beam search decoder for phrase-based statistical machine models, user manual and description for version 1.2*, 2004.

Testset	System	BLEU score	NIST score
devset3	baseline w/o MERT	0.3247	6.9353
	our method w/o MERT	0.3881	7.9094
reordered devset3	our method w/o MERT	0.3675	7.6214
ASR 1-best	baseline w/o MERT	0.1081	4.3555
	our method w/o MERT	0.1366	4.8438
	our method w/ MERT	0.1311	4.8372
correct transcription	baseline w/o MERT	0.1170	4.7078
	our method w/o MERT	0.1459	5.3649
	our method w/ MERT	0.1431	5.2105

Table 2: Translation accuracy for devset3, ASR 1-best recognition result, and correct transcription (all systems used devset3 for training)

[9] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[10] S. Kurohashi and M. Nagao, "Building a Japanese Parsed Corpus while Improving the Parsing System," in *Proceedings of the NLPRS*, 1997, pp. 451–456.

[11] M. Komachi, R. Iida, K. Inui, and Y. Matsumoto, "Verbal Noun Argument Structure Analysis Based on Co-occurrence and Patterns of Noun Phrases," in *Proceedings of the 12th Annual Meeting of the Association for NLP (in Japanese)*, 2006, pp. 821–824.

[12] R. Iida, K. Inui, and Y. Matsumoto, "Exploiting Syntactic Patterns as Clues in Zero-Anaphora Resolution," in *Proceedings of the 21st International Conference on Computational Linguistics (COLING) and 44th Annual Meeting of the ACL*, 2006, pp. 625–632.

[13] H. Suzuki and K. Toutanova, "Learning to Predict Case Markers in Japanese," in *Proceedings of the 21st International Conference on Computational Linguistics (COLING) and 44th Annual Meeting of the ACL*, 2006, pp. 1049–1056.

[14] M. Palmer, D. Gildea, and P. Kingsbury, "The Proposition Bank: A Corpus Annotated with Semantic Roles," *Computational Linguistics*, vol. 31, no. 1, 2005.

[15] N. Xue and M. Palmer, "Annotating Propositions in the Penn Chinese Treebank," in *Proceedings of the Second Sighan Workshop*, 2003.

[16] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara, *Morphological Analysis Aystem ChaSen, ver. 2.3.3*, http://chasen.naist.jp/, 2003.

[17] T. Kudo and Y. Matsumoto, "Japanese Dependency Analysis using Cascaded Chunking," in *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)*, 2002, pp. 63–69.

[18] A. Ratnaparkhi, "A Maximum Entropy Part-Of-Speech Tagger," in *Proceedings of the Empirical Methods in Natural Language Processing Conference*, 1996.

[19] A. Ito, "Palmkit," http://palmkit.sourceforge.net/, 2002.

[20] A. Venugopal, "Minimum Error Rate Training Tool," http://www.cs.cmu.edu/˜ashishv/mer.html, 2005.