Improving Statistical Machine Translation with Processing Shallow Parsing

Hoai-Thu Vuong, Vinh Van Nguyen
University of Engineering and Technology,
Vietnam National University
144, Xuan Thuy, Cau Giay , Hanoi
{thuvh, vinhnv}@vnu.edu.vn

Viet Hong Tran
Department of Information Technology
University Of Economic And Technical Industries
456 Minh Khai, Hai Ba Trung, Hanoi
thviet@uneti.edu.vn

Akira Shimazu
Japan Advanced Institute of Science and Technology
1-1, Asahidai, Nomi, Ishikawa, 923-1292, Japan
shimazu@jaist.ac.jp

Abstract

Reordering is of essential importance for phrase based statistical machine translation (SMT). In this paper, we would like to present a new method of reordering in phrase based SMT. We inspired from (Xia and McCord, 2004) using preprocessing reordering approaches. We used shallow parsing and transformation rules to reorder the source sentence. The experiment results from English-Vietnamese pair showed that our approach achieves significant improvements over MOSES which is the state-of-the art phrase based system.

1 Introduction

In SMT, the reordering problem (global reordering) is one of the major problems, since different languages have different word order requirements. The SMT task can be viewed as two subtasks: predicting the collection of words in a translation, and deciding the order of the predicted words (reordering problem). Currently, phrase-based statistical machine translation (Koehn et al., 2003; Och and Ney, 2004) is the state-of-the-art of SMT because of its power in modelling short reordering and local context.

However, with phrase based SMT, long distance reordering is still problematic. In order to tackle the long distance reordering problem, in recent years, huge research efforts have been conducted using syntactic information. There are some studies on integrating syntactic resources within SMT. Chiang (Chiang, 2005) shows significant improvement by keeping the strengths of phrases, while incorporating syntax into SMT. Some approaches have been applied at the word-level (Collins et al., 2005). They are particularly useful for language with rich morphology, for reducing data sparseness. Other kinds of syntax reordering methods require parser trees , such as the work in (Quirk et al., 2005; Collins et al., 2005; Huang and Mi, 2010). The parsed tree is more powerful in capturing the sentence structure. However, it is expensive to create tree structure, and building a good quality parser is also a hard task. All the above approaches require much decoding time, which is expensive.

The approach we are interested in here is to balance the quality of translation with decoding time. Reordering approaches as a preprocessing step (Xia and McCord, 2004; Xu et al., 2009; Talbot et al., 2011; Katz-Brown et al., 2011) is very effective (improvement significant over state of-the-art phrase-based and hierarchical machine translation systems and separately quality evaluation of reordering models).

Inspiring this preprocessing approach, we have proposed a combine approach which preserves the strength of phrase-based SMT in local reordering and decoding time as well as the strength of integrating syntax in reordering. Consequently, we use an intermediate syntax between POS tag and parse tree: shallow parsing. Firstly, we use shallow parsing for preprocessing with training and testing. Second, we apply a series of transformation rules which are learnt automatically from parallel corpus to the shallow tree. The experiment results from English-Vietnamese pair showed that our approach achieves

26th Pacific Asia Conference on Language,Information and Computation pages 401–407

significant improvements over MOSES which is the state-of-the art phrase based system.

The rest of this paper is structured as follows. Section 2 reviews the related works. Section 3 briefly introduces phrase-based SMT. Section 4 introduces how to apply transformation rules to the shallow tree. Section 5 describes and discusses the experimental results. And, conclusions are given in Section 6.

2 Related works

As mentioned in section 1, some approaches using syntactic information are applied to solve the reordering problem. One of approaches is syntactic parsing of source language and reordering rules as preprocessing steps. The main idea is transferring the source sentences to get very close target sentences in word order as possible, so EM training is much easier and word alignment quality becomes better. There are several studies to improve reordering problem such as (Xia and McCord, 2004; Collins et al., 2005; Nguyen and Shimazu, 2006; Wang et al., 2007; Habash, 2007; Xu et al., 2009).

They all performed reordering during preprocessing step based on the source tree parsing combining either automatic extracted syntactic rules (Xia and McCord, 2004; Nguyen and Shimazu, 2006; Habash, 2007) or handwritten rules (Collins et al., 2005; Wang et al., 2007; Xu et al., 2009).

(Xu et al., 2009) described method using dependency parse tree and a flexible rule to perform the reordering of subject, object, etc... These rules were written by hand, but (Xu et al., 2009) showed that an automatic rule learner can be used.

(Collins et al., 2005) developed a clause detection and used some handwritten rules to reorder words in the clause. Partly, (Xia and McCord, 2004; Habash, 2007) built an automatic extracted syntactic rules.

Compared with theses approaches, our work has a few differences. Firstly, we aim to develop the phrase-based translation model to translate from English to Vietnamese. Secondly, we build a shallow tree by chunking in recursively (chunk of chunk). Thirdly, we use not only the automatic rules, but also some handwritten rules, to transform the source sentence. As the same with (Xia and McCord, 2004; Habash, 2007), we also apply preprocessing in both training and decoding time.

The other approaches use syntactic parsing to provide multiple source sentence reordering options through word (phrase) lattices (Zhang et al., 2007; Nguyen et al., 2007). (Nguyen et al., 2007) applied some transformation rules, which is learnt automatically from bilingual corpus, to reorder some words in a chunk. A crucial difference between their methods and ours is that they do not perform reordering during training. While, our method can solve this problem by using a complicated structure, which is more efficient with a shallow tree (chunk of chunks).

3 Brief description of the baseline Phrase-based SMT

In this section, we will describe the phrase-based SMT system which was used for the experiments. Phrase-based SMT, as described by (Koehn et al., 2003) translates a source sentence into a target sentence by decomposing the source sentence into a sequence of source phrases, which can be any contiguous sequences of words (or tokens treated as words) in the source sentence. For each source phrase, a target phrase translation is selected, and the target phrases are arranged in some order to produce the target sentence. A set of possible translation candidates created in this way is scored according to a weighted linear combination of feature values, and the highest scoring translation candidate is selected as the translation of the source sentence. Symbolically,

$$\hat{t} = \arg\max_{t,a} \sum_{i=1}^{n} \lambda_i f_j(s,t,a) \qquad (1)$$

when s is the input sentence, t is a possible output sentence, and a is a phrasal alignment that specifies how t is constructed from s, and $\hat{t}$ is the selected output sentence. The weights λ_i associated with each feature f_i are tuned to maximize the quality of the translation hypothesis selected by the decoding procedure that computes the argmax. The log-linear model is a natural framework to integrate many features. The baseline system uses the following features:

- the probability of each source phrase in the hypothesis given the corresponding target phrase.

- the probability of each target phrase in the hypothesis given the corresponding source phrase.
- the lexical score for each target phrase given the corresponding source phrase.
- the lexical score for each source phrase given the corresponding target phrase.
- the target language model probability for the sequence of target phrase in the hypothesis.
- the word and phrase penalty score, which allow to ensure that the translation does not get too long or too short.
- the distortion model allows for reordering of the source sentence.

The probabilities of source phrase given target phrases, and target phrases given source phrases, are estimated from the bilingual corpus.

(Koehn et al., 2003) used the following distortion model (reordering model), which simply penalizes nonmonotonic phrase alignment based on the word distance of successively translated source phrases with an appropriate value for the parameter α:

$$d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1} - 1|} \quad (2)$$

4 Shallow Syntactic Preprocessing for SMT

In this section, we describe the transformation rules and how applying it to shallow tree for reordering an English sentence.

4.1 Transformation Rule

Suppose that T_s is a given lexicalized tree of the source language (whose nodes are augmented to include a word and a POS label). T_s contains n applications of lexicalized CFG rules $LHS_i \rightarrow RHS_i$ ($i \in \overline{1,n}$). We want to transform T_s into the target language word order by applying transformational rules to the CFG rules. A transformational rule is represented as ($LHS \rightarrow RHS, RS$), which is a pair consisting of an unlexicalized CFG rule and a reordering sequence (RS). For example, the rule (NP $\rightarrow$ JJ NN, 1 0) implies that the CFG rule (NP$\rightarrow$ JJ NN) in the source language can be transformed into the rule (NP$\rightarrow$NN JJ) in the target language. Since the possible transformational rule for each CFG rule is not unique, there can be many transformed trees. The problem is how to choose the best one (we can see (Nguyen and Shimazu, 2006) for a description in more details). We use the method described in (Nguyen and Shimazu, 2006) to extract the transformation rules from the parallel corpus and induce the best sequence of transformation rules for a source tree. Besides, we also built a small set of transformation rules by hand (the handwritten rules).

4.2 Shallow Syntactic Processing

In this section, we describe a method to build a translation model for a pair English to Vietnamese. We aim to reorder an English sentence to get a new English, and some words in this sentence are arranged as Vietnamese words order.

tom	's	[two	blue	books]
tom	's	[two	books	blue]
[two	books	blue]	's	tom
hai	cuốn sách	màu xanh	của	tom

Figure 1: An Example of phrase before and after our preprocessing

Figure 1 gives examples of original and preprocessed phrase in English. The first line is the original English phrase with a chunk (two blue books), and the second line is the phrase with a modified chunk (two books blue). This chunk is arranged as the Vietnamese order. However, we aim to preprocess the words outside the chunk (the phrase "tom 's" in Figure 1), and the third line is the output of our method. Finally, the fourth line is the Vietnamese phrase. As you can see, the third and fourth line have the same word order.

After pre-processed, this new sentence is used in training process to get a phrased translation model, and in decoding process to get a target sentence (by using translation which is trained in training process). To preprocess, we follow these steps:

- building shallow syntactic

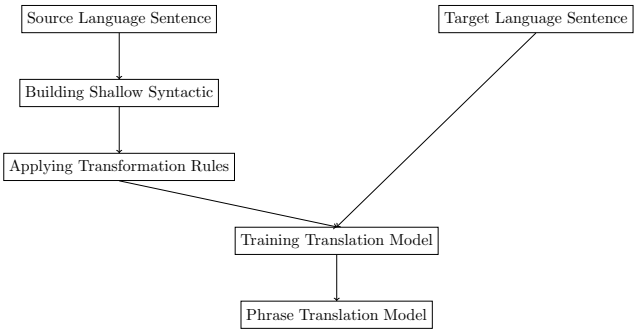

Figure 2: Our training process

- applying transformation rules

So as to build shallow syntactic, we use a method described in (Tsuruoka et al., 2009). Their approach introduced the method to parse an English sentence by using chunking (balance accuracy with speed time). Their method is high accuracy (accuracy with 88.4 F-score) and fast parsing time: using CRFTagger to chunk the sentence, and then setup a tree from the chunks and recursive until they cannot chunk the sentence. Their result showed that this method is outstanding in performance with high accuracy. As they did, we also receive a shallow syntactic when parse the source sentence in English. However, we stop chunking after two loop steps. So that, *the highest deep of node in syntactic tree is two*. By doing that, we will balance between accuracy and performance time. We can use the method of (Tsuruoka et al., 2009) to build full parse tree, but that will be leave it for future work.

After building the shallow syntactic, the transformation rules are applied. After finding the matching rule from the top of the shallow tree, we arrange the words in the English sentence, which is covered by the matching node, like Vietnamese words order. And then, we do the same for each children of this node. If any rule is applied, we use the order of original sentence. Not only rule is learnt automatically from bilingual corpora, we also try applying handwritten rules.

5 Experiment

5.1 Implementation

- We developed the shallow parsing by using the method from (Tsuruoka et al., 2009) to parse a source sentences (English sentences) including a shallow tree.

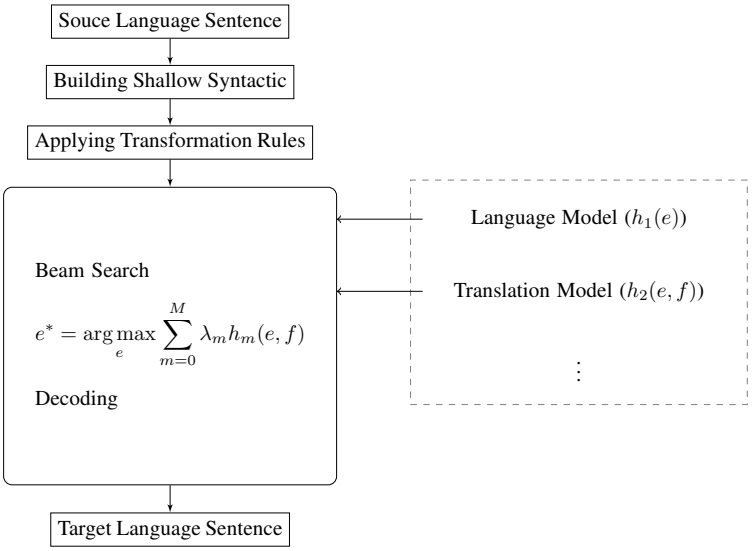

Figure 3: Our decoding process

- The rules are learnt from English-Vietnamese parallel corpus and Penntree Bank Corpus. We used the CFG transformation rules (chunk levels) for extraction from (Nguyen and Shimazu, 2006)'s method to reorder shallow tree of a source sentences.
- We implemented preprocessing step during both training and decoding time.
- Using the SMT Moses decoder (Koehn et al., 2007) for decoding.

5.2 Data set and Experimental Setup

For evaluation, we used an English-Vietnamese corpus (Nguyen et al., 2008), including about 54642 pairs for training, 500 pairs for testing and 200 pairs for development test set. Table 1 gives more statistical information about our corpora. We conducted some experiments with SMT Moses Decoder (Koehn et al., 2007) and SRILM (Stolcke, 2002). We trained a trigram language model using interpolate and kndiscount smoothing with 89M Vietnamese mono corpus. Before extracting phrase table, we use GIZA++ (Och and Ney, 2003) to build word alignment with grow-diag-final-and algorithm.

Corpus	Sentence pairs	Training Set	Development Set	Test Set
General	55341	54642	200	499

		English	Vietnamese
Training	Sentences	54620	
	Average Length	11.2	10.6
	Word	614578	580754
	Vocabulary	23804	24097
Development	Sentences	200	
	Average Length	11.1	10.7
	Word	2221	2141
	Vocabulary	825	831
Test	Sentences	499	
	Average Length	11.2	10.5
	Word	5620	6240
	Vocabulary	1844	1851

Table 1: Corpus Statistical

Besides using preprocessing, we also used default reordering model in Moses Decoder: using word-based extraction (wbe), splitting type of reodering orientation to three class (monotone, swap and discontinuous – msd), combining backward and forward direction (bidirectional) and modeling base on both source and target language (fe) (Koehn et al., 2007). First system in 2 is our baseline system. The second and the third system are the baseline system which is applied the transformation rules (include the automatic and handwritten rules). In these experiments, we only use the chunking level. The fifth experiment is the result of our works: applied automatic transformation rules into shallow syntactic. By doing these experiments, we can show the effective of our method. In addition, we also did the fourth and sixth experiment with a specific parameter for the MOSES Decoder (monotone). By using this flag, we will discard the distortion model, so that, the decoder only do monotone decode.

5.3 BLEU score

The result of our experiments in table 3 showed our applying transformation rule to process the source sentences. Thanks to this method, we can find out various phrases in the translation model. So that, they enable us to have more options for decoder to generate the best translation.

Table 4 describes the BLEU score (Papineni et al., 2002) of our experiments. As we can see, by applying preprocess in both training and decoding, the BLEU score of our best system increase by 0.82 point "Baseline + AR (shallow syntactic)" system) over "Baseline system". Improvement over 0.82 BLEU point is valuable because baseline system is the strong phrase based SMT (integrating lexicalized reordering models). The improvement of "Baseline + AR (shallow syntactic)" system is statistically significant at $p < 0.01$.

System	BLEU (%)
Baseline	36.84
Baseline + MR	37.33
Baseline + AR	37.24
Baseline + AR (monotone)	35.80
Baseline + AR (shallow syntactic)	**37.66**
Baseline + AR (shallow syntactic + monotone)	37.43

Table 4: Translation performance for the English-Vietnamese task

We also carried out the experiments with handwritten rules. Using some handwritten rules help the phrased translation model generate some best translation more than the automatic rules. Besides, the result proved that the effect of applying transformation rule on the shallow syntactic when the BLEU score is highest. Because, the cover of handwritten rules is larger than the automatic rules.

Furthermore, handwritten rule is made by human, and focus on popular cases. So that, we get some

Name	Description
Baseline	Phrase-based system
Baseline + MR	Phrase-based system with corpus which is preprocessed using handwritten rules
Baseline + AR	Phrase-based system with corpus which is preprocessed using automatic learning rules
Baseline + AR (monotone)	Phrase-based system with corpus which is preprocessed using automatic learning rules and decoded by monotone decoder
Baseline + AR(shallow syntactic)	Phrase-based system with corpus which is shallow syntactic analyze and applied automatic transformation rules
Baseline + AR(shallow syntactic+monotone)	Phrase-based system with corpus which is shallow syntactic analyze and applied automatic transformation rules

Table 2: Details of our experimental, AR is named as using automatic rules, MR is named as using handwritten rules

Name	Size of phrase-table
Baseline	1237568
Baseline + MR	1251623
Baseline + AR	1243699
Baseline + AR (monotone)	1243699
Baseline + AR (shallow syntactic)	**1279344**
Baseline + AR (shallow syntactic + monotone)	1279344

Table 3: Size of phrase tables

pair of sentences with the best alignment, and then, we can extract more and better phrase tables. Finally, the BLEU score of using monotone decoder decrease by 1% when we use preprocessing in only base chunk level, and our shallow syntactic decreased a bit. As, the default reordering model in baseline system is better than in this experiment[1].

6 Conclusion

In this paper, we would like to present a new method for reordering in phrase based SMT. We inspired from (Xia and McCord, 2004) using preprocessing reordering approaches. We used shallow parsing and transformation rules for reordering the source sentence. Meanwhile, we limit the height of syntactic tree to balance the accuracy with performance of system. The experiment results with English-Vietnamese pair showed that our approach achieves significant improvements over MOSES which is the state-of-the art phrase based system. In the future, we would like to evaluate our method with tree with higher and deeper syntactic structure and larger size of corpus.

[1]The reordering model in the monotone decoder is distance based, introduced in (Koehn et al., 2003). This model is a default reordering model in Moses Decoder (Koehn et al., 2007)

Acknowledgment

This work described in this paper has been partially funded by Hanoi National University (QG.12.49 project) and the Vietnam National Foundation for Science and Technology Development (Nafosted).

References

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June.

M. Collins, P. Koehn, and I. Kucerová. 2005. Clause restructuring for statistical machine translation. In *Proc. ACL 2005*, pages 531–540. Ann Arbor, USA.

N. Habash. 2007. Syntactic preprocessing for statistical machine translation. *Proceedings of the 11th MT Summit*.

Liang Huang and Haitao Mi. 2010. Efficient incremental decoding for tree-to-string translation. In *Proceedings*

of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 273–283, Cambridge, MA, October. Association for Computational Linguistics.

Jason Katz-Brown, Slav Petrov, Ryan McDonald, Franz Och, David Talbot, Hiroshi Ichikawa, Masakazu Seno, and Hideto Kazawa. 2011. Training a parser for machine translation reordering. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 183–192, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 127–133. Edmonton, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL, Demonstration Session*.

Thai Phuong Nguyen and Akira Shimazu. 2006. Improving phrase-based smt with morpho-syntactic analysis and transformation. In *Proceedings AMTA 2006*.

Phuong Thai Nguyen, Akira Shimazu, Le-Minh Nguyen, and Van-Vinh Nguyen. 2007. A syntactic transformation model for statistical machine translation. *International Journal of Computer Processing of Oriental Languages (IJCPOL)*, 20(2):1–20.

Thai Phuong Nguyen, Akira Shimazu, Tu Bao Ho, Minh Le Nguyen, and Vinh Van Nguyen. 2008. A tree-to-string phrase-based model for statistical machine translation. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL 2008)*, pages 143–150, Manchester, England, August. Coling 2008 Organizing Committee.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz J. Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318. Philadelphia, PA, July.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal smt. In *Proceedings of ACL 2005*, pages 271–279. Ann Arbor, Michigan, USA.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, volume 29, pages 901–904.

David Talbot, Hideto Kazawa, Hiroshi Ichikawa, Jason Katz-Brown, Masakazu Seno, and Franz Och. 2011. A lightweight evaluation framework for machine translation reordering. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 12–21, Edinburgh, Scotland, July. Association for Computational Linguistics.

Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2009. Fast full parsing by linear-chain conditional random fields. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 790–798, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 737–745, Prague, Czech Republic, June. Association for Computational Linguistics.

Fei Xia and Michael McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of Coling 2004*, pages 508–514, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. 2009. Using a dependency parser to improve smt for subject-object-verb languages. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 245–253, Boulder, Colorado, June. Association for Computational Linguistics.

Yuqi Zhang, Richard Zens, and Hermann Ney. 2007. Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 1–8.