Syntactic Constraints on Phrase Extraction for Phrase-Based Machine Translation

Hailong Cao, Andrew Finch and Eiichiro Sumita
Language Translation Group, MASTAR Project
National Institute of Information and Communications Technology
{hlcao,andrew.finch,eiichiro.sumita }@nict.go.jp

Abstract

A typical phrase-based machine translation (PBMT) system uses phrase pairs extracted from word-aligned parallel corpora. All phrase pairs that are consistent with word alignments are collected. The resulting phrase table is very large and includes many non-syntactic phrases which may not be necessary. We propose to filter the phrase table based on source language syntactic constraints. Rather than filter out all non-syntactic phrases, we only apply syntactic constraints when there is phrase segmentation ambiguity arising from unaligned words. Our method is very simple and yields a 24.38% phrase pair reduction and a 0.52 BLEU point improvement when compared to a baseline PBMT system with full-size tables.

1 Introduction

Both PBMT models (Koehn et al., 2003; Chiang, 2005) and syntax-based machine translation models (Yamada et al., 2000; Quirk et al., 2005; Galley et al., 2006; Liu et al., 2006; Marcu et al., 2006; and numerous others) are the state-of-the-art statistical machine translation (SMT) methods. Over the last several years, an increasing amount of work has been done to combine the advantages of the two approaches. DeNeefe et al. (2007) made a quantitative comparison of the phrase pairs that each model has to work with and found it is useful to improve the phrasal coverage of their string-to-tree model. Liu et al. (2007) proposed forest-to-string rules to capture the non-syntactic phrases in their tree-to-string model. Zhang et al. (2008) proposed a tree sequence based tree-to-tree model which can describe non-syntactic phrases with syntactic structure information.

The converse of the above methods is to incorporate syntactic information into the PBMT model. Zollmann and Venugopal (2006) started with a complete set of phrases as extracted by traditional PBMT heuristics, and then annotated the target side of each phrasal entry with the label of the constituent node in the target-side parse tree that subsumes the span. Marton and Resnik (2008) and Cherry (2008) imposed syntactic constraints on the PBMT system by making use of prior linguistic knowledge in the form of syntax analysis. In their PBMT decoders, a candidate translation gets an extra credit if it respects the source side syntactic parse tree but may incur a cost if it violates a constituent boundary. Xiong et al. (2009) proposed a syntax-driven bracketing model to predict whether a phrase (a sequence of contiguous words) is bracketable or not using rich syntactic constraints.

In this paper, we try to utilize syntactic knowledge to constrain the phrase extraction from word-based alignments for PBMT system. Rather than filter out all non-syntactic phrases, we only apply syntactic constraints when there is phrase segmentation ambiguity arising from unaligned words. Our method is very simple and yields a 24.38% phrase pair reduction and a 0.52 BLEU point improvement when compared to the baseline PBMT system with full-size tables.

2 Extracting Phrase Pairs from Word-based Alignments

In this section, we briefly review a simple and effective phrase pair extraction algorithm upon which this work builds.

Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation, pages 28–33,
COLING 2010, Beijing, August 2010.

The basic translation unit of a PBMT model is the phrase pair, which consists of a sequence of source words, a sequence of target words and a vector of feature values which represents this pair's contribution to the translation model. In typical PBMT systems such as MOSES (Koehn, 2007), phrase pairs are extracted from word-aligned parallel corpora. Figure 1 shows the form of training example.

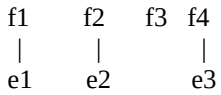

Figure 1: An example parallel sentence pair and word alignment

Since there is no phrase segmentation information in the word-aligned sentence pair, in practice all pairs of "source word sequence ||| target word sequence" that are consistent with word alignments are collected. The words in a legal phrase pair are only aligned to each other, and not to words outside (Och et al., 1999). For example, given a sentence pair and its word alignments shown in Figure1, the following nine phrase pairs will be extracted:

Source phrase ||| Target phrase
f1 ||| e1
f2 ||| e2
f4 ||| e3
f1 f2 ||| e1 e2
f2 f3 ||| e2
f3 f4 ||| e3
f1 f2 f3 ||| e1 e2
f2 f3 f4 ||| e2 e3
f1 f2 f3 f4 ||| e1 e2 e3

Table 1: Phrase pairs extracted from the example in Figure 1

Note that neither the source phrase nor the target phrase can be empty. So "f3 ||| EMPTY" is not a legal phrase pair.

Phrase pairs are extracted over the entire training corpus. Given all the collected phrase pairs, we can estimate the phrase translation probability distribution by relative frequency. The collected phrase pairs will also be used to build the lexicalized reordering model. For more details of the lexicalized reordering model, please refer to Tillmann and Zhang (2005) and section 2.7.2 of the MOSES's manual[1].

The main problem of such a phrase pair extraction procedure is the resulting phrase translation table is very large, especially when a large quantity of parallel data is available. This is not desirable in real application where speed and memory consumption are often critical concerns. In addition, some phrase translation pairs are generated from training data errors and word alignment noise. Therefore, we need to filter the phrase table in an appropriate way for both efficiency and translation quality (Johnson et al., 2007; Yang and Zheng, 2009).

3 Syntactic Constraints on Phrase Pair Extraction

We can divide all the possible phrases into two types: syntactic phrases and non-syntactic phrases. A "syntactic phrase" is defined as a word sequence that is covered by a single subtree in a syntactic parse tree (Imamura, 2002). Intuitively, we would think syntactic phrases are much more reliable while the non-syntactic phrases are useless. However, (Koehn et al., 2003) showed that restricting phrasal translation to only syntactic phrases yields poor translation performance – the ability to translate non-syntactic phrases (such as "there are", "note that", and "according to") turns out to be critical and pervasive.

(Koehn et al., 2003) uses syntactic constraints from both the source and target languages, and over 80% of all phrase pairs are eliminated. In this section, we try to use syntactic knowledge in a less restrictive way.

Firstly, instead of using syntactic restriction on both source phrases and target phrases, we only apply syntactic restriction to the source language side.

Secondly, we only apply syntactic restriction to the source phrase whose first or last word is unaligned.

For example, given a parse tree illustrated in Figure 2, we will filter out the phrase pair "f2 f3 ||| e2" since the source phrase "f2 f3" is a non-syntactic phrase and its last word "f3" is not

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

aligned to any target word. The phrase pair "f1 f2 f3 ||| e1 e2" will also be eliminated for the same reason. But we do keep phrase pairs such as "f1 f2 ||| e1 e2" even if its source phrase "f1 f2" is a non-syntactic phrase. Also, we keep "f3 f4 ||| e3" since "f3 f4" is a syntactic phrase. Table 2 shows the completed set of phrase pairs that are extracted with our constraint-based method.

Source phrase ||| Target phrase
f1 ||| e1
f2 ||| e2
f4 ||| e3
f1 f2 ||| e1 e2
f3 f4 ||| e3
f2 f3 f4 ||| e2 e3
f1 f2 f3 f4 ||| e1 e2 e3

Table 2: Phrase pairs extracted from the example in Figure 2

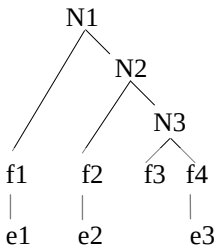

Figure 2: An example parse tree and word-based alignments

The state-of-the-art alignment tool such as GIZA++[2] can not always find alignments for every word in the sentence pair. The possible reasons could be: its frequency is too low, noisy data, auxiliary words or function words which have no obvious correspondence in the opposite language.

In the automatically aligned parallel corpus, unaligned words are frequent enough to be noticeable (see section 4.1 in this paper). How to decide the translation of unaligned word is left to the phrase extraction algorithm. An unaligned source word should be translated together with the words on the right of it or the words on the left of it. The existing algorithm considers both of the two directions. So both "f2 f3 ||| e2" and "f3 f4 ||| e3" are extracted. However, it is unlikely that "f3" can be translated into both "e2" and "e3". So our algorithm uses prior syntactic knowledge to keep "f3 f4 ||| e3" and exclude "f2 f3 ||| e2".

4 Experiments

Our SMT system is based on a fairly typical phrase-based model (Finch and Sumita, 2008). For the training of our SMT model, we use a modified training toolkit adapted from the MOSES decoder. Our decoder can operate on the same principles as the MOSES decoder. Minimum error rate training (MERT) with respect to BLEU score is used to tune the decoder's parameters, and it is performed using the standard technique of Och (2003). A lexicalized reordering model was built by using the "msd-bidirectional-fe" configuration in our experiments.

The translation model was created from the FBIS parallel corpus. We used a 5-gram language model trained with modified Kneser-Ney smoothing. The language model was trained on the target side of the FBIS corpus and the Xinhua news in the GIGAWORD corpus. The development and test sets are from the NIST MT08 evaluation campaign. Table 3 shows the statistics of the corpora used in our experiments.

Data	Sentences	Chinese words	English words
Training set	221,994	6,251,554	8,065,629
Development set	1,664	38,779	46,387
Test set	1,357	32,377	42,444
GIGAWORD	19,049,757	-	306,221,306

Table 3: Corpora statistics

The Chinese sentences are segmented, POS tagged and parsed by the tools described in Kruengkrai et al. (2009) and Cao et al. (2007), both of which are trained on the Penn Chinese Treebank 6.0.

[2] http://fjoch.com/GIZA++.html

4.1 Experiments on Word Alignments

We use GIZA++ to align the sentences in both the Chinese-English and English-Chinese directions. Then we combine the alignments using the standard "grow-diag-final-and" procedure provided with MOSES.

In the combined word alignments, 614,369 or 9.82% of the Chinese words are unaligned. Table 4 shows the top 10 most frequently unaligned words. Basically, these words are auxiliary words or function words whose usage is very flexible. So it would be difficult to automatically align them to the target words.

Unaligned word	Frequency
的	77776
,	29051
在	9414
一	8768
中	8543
个	7471
是	7365
上	6155
了	5945
不	5450

Table 4: Frequently unaligned words from the training corpus

4.2 Experiments on Chinese-English SMT

In order to confirm that it is advantageous to apply appropriate syntactic constraints on phrase extraction, we performed three translation experiments by using different ways of phrase extraction.

In the first experiment, we used the method introduced in Section 2 to extract all possible phrase translation pairs without using any constraints arising from knowledge of syntax.

The second experiment used source language syntactic constraints to filter out all non-syntactic phrases during phrase pair extraction.

The third experiment used source language syntactic constraints to filter out only non-syntactic phrases whose first or last source word was unaligned.

With the exception of the above differences in phrase translation pair extraction, all the other settings were the identical in the three experiments. Table 5 summarizes the SMT performance. The evaluation metric is case-sensitive BLEU-4 (Papineni et al., 2002) which estimates the accuracy of translation output with respect to a set of reference translations.

Syntactic Constraints	Number of distinct phrase pairs	BLEU
None	14,195,686	17.26
Full constraint	4,855,108	16.51
Selectively constraint	10,733,731	17.78

Table 5: Comparison of different constraints on phrase pair extraction by translation quality

As shown in the table, it is harmful to fully apply syntactic constraints on phrase extraction, even just on the source language side. This is consistent with the observation of (Koehn et al., 2003) who applied both source and target constraints in German to English translation experiments.

Clearly, we obtained the best performance if we use source language syntactic constraints only on phrases whose first or last source word is unaligned. In addition, we reduced the number of distinct phrase pairs by 24.38% over the baseline full-size phrase table.

The results in table 5 show that while some non-syntactic phrases are very important to maintain the performance of a PBMT system, not all of them are necessary. We can achieve better performance and a smaller phrase table by applying syntactic constraints when there is phrase segmentation ambiguity arising from unaligned words.

5 Related Work

To some extent, our idea is similar to Ma et al. (2008), who used an anchor word alignment model to find a set of high-precision anchor links and then aligned the remaining words relying on dependency information invoked by the acquired anchor links. The similarity is that both Ma et al. (2008) and this work utilize structure information to find appropriate translations for words which are difficult to align. The differ-

ence is that they used dependency information in the word alignment stage while our method uses syntactic information during the phrase pair extraction stage. There are also many works which leverage syntax information to improve word alignments (e.g., Cherry and Lin, 2006; DeNero and Klein, 2007; Fossum et al., 2008; Hermjakob, 2009).

Johnson et al., (2007) presented a technique for pruning the phrase table in a PBMT system using Fisher's exact test. They compute the significance value of each phrase pair and prune the table by deleting phrase pairs with significance values smaller than a certain threshold. Yang and Zheng (2008) extended the work in Johnson et al., (2007) to a hierarchical PBMT model, which is built on synchronous context free grammars (SCFG). Tomeh et al., (2009) described an approach for filtering phrase tables in a statistical machine translation system, which relies on a statistical independence measure called *Noise*, first introduced in (Moore, 2004). The difference between the above research and this work is they took advantage of some statistical measures while we use syntactic knowledge to filter phrase tables.

6 Conclusion and Future Work

Phrase pair extraction plays a very important role on the performance of PBMT systems. We utilize syntactic knowledge to constrain the phrase extraction from word-based alignments for a PBMT system. Rather than filter out all non-syntactic phrases, we only filter out non-syntactic phrases whose first or last source word is unaligned. Our method is very simple and yields a 24.38% phrase pair reduction and a 0.52 BLEU point improvement when compared to the baseline PBMT system with full-size tables.

In the future work, we will use other language pairs to test our phrase extraction method so that we can discover whether or not it is language independent.

References

Robert C. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *EMNLP*.

Hailong Cao, Yujie Zhang and Hitoshi Isahara. Empirical study on parsing Chinese based on Collins' model. 2007. In *PACLING*.

Colin Cherry and Dekang Lin. 2006. Soft syntactic constraints for word alignment through discriminative training. In *ACL*.

Colin Cherry. 2008. Cohesive phrase-Based decoding for statistical machine translation. In *ACL-HLT*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *ACL*.

Steve DeNeefe, Kevin Knight, Wei Wang, and Daniel Marcu. 2007. What can syntax-based MT learn from phrase-based MT? In *EMNLP-CoNLL*.

John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *ACL*.

Andrew Finch and Eiichiro Sumita. 2008. Dynamic model interpolation for statistical machine translation. In *SMT Workshop*.

Victoria Fossum, Kevin Knight and Steven Abney. 2008. Using syntax to improve word alignment precision for syntax-based machine translation. In *SMT Workshop, ACL*.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve Deneefe, Wei Wang and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *ACL*.

Ulf Hermjakob. 2009. Improved word alignment with statistics and linguistic heuristics. In *EMNLP*.

Kenji Imamura. 2002. Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT. In *TMI*.

Howard Johnson, Joel Martin, George Foster and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrase table. In *EMNLP-CoNLL*.

Franz Josef Och, Christoph Tillmann and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *EMNLP-VLC*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL*.

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, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL demo and poster sessions.*

Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa and Hitoshi Isahara. 2009. An error-driven word-character hybrid model for joint Chinese word segmentation and POS tagging. In *ACL-IJCNLP*.

Yang Liu, Qun Liu, Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *ACL-COLING*.

Yang Liu, Yun Huang, Qun Liu and Shouxun Lin. 2007. Forest-to-string statistical translation rules. In *ACL*.

Yanjun Ma, Sylwia Ozdowska, Yanli Sun and Andy Way. 2008. Improving word alignment using syntactic dependencies. In *SSST*.

Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. 2006. SPMT: Statistical machine translation with syntactified target language phrases. In *EMNLP*.

Yuval Marton and Philip Resnik. 2008. Soft syntactic constraints for hierarchical phrased-based translation. In *ACL-HLT*.

Kishore Papineni, Salim Roukos, Todd Ward and WeiJing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Chris Quirk and Arul Menezes and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *ACL*.

Christoph Tillmann and Tong Zhang. 2005. A localized prediction model for statistical machine translation. In *ACL*.

Nadi Tomeh, Nicola Cancedda and Marc Dymetman. 2009. Complexity-based phrase-table filtering for statistical machine translation. In *MT Summit*.

Deyi Xiong, Min Zhang, Aiti Aw and Haizhou Li. 2009. A syntax-driven bracketing model for phrase-based translation. In *ACL-IJCNLP*.

Kenji Yamada and Kevin Knight. 2000. A syntax-based statistical translation model. In *ACL*.

Mei Yang and Jing Zheng. 2009. Toward smaller, faster, and better hierarchical phrase-based SMT. In *ACL*.

Min Zhang, Hongfei Jiang, Aiti Aw, Chew Lim Tan and Sheng Li. 2008. A tree sequence alignment-based tree-to-tree translation model. In *ACL- HLT*.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *SMT Workshop, HLT-NAACL*.