Phrasal Syntactic Category Sequence Model for Phrase-Based MT

Hailong Cao[1], Eiichiro Sumita[2], Tiejun Zhao[1], and Sheng Li[1]

[1] Harbin Institute of Technology, China
[2] National Institute of Information and Communications Technology, Japan
{hailong,tjzhao,shengli}@mtlab.hit.edu.cn,
eiichiro.sumita@nict.go.jp

Abstract. Incorporating target syntax into phrase-based machine translation (PBMT) can generate syntactically well-formed translations. We propose a novel phrasal syntactic category sequence (PSCS) model which allows a PBMT decoder to prefer more grammatical translations. We parse all the sentences on the target side of the bilingual training corpus. In the standard phrase pair extraction procedure, we assign a syntactic category to each phrase pair and build a PSCS model from the parallel training data. Then, we log linearly incorporate the PSCS model into a standard PBMT system. Our method is very simple and yields a 0.7 BLEU point improvement when compared to the baseline PBMT system.

Keywords: machine translation, natural language processing, phrase-based machine translation.

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 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

A. Gelbukh (Ed.): CICLing 2012, Part II, LNCS 7182, pp. 52–59, 2012.
© Springer-Verlag Berlin Heidelberg 2012

subsumes the span. Hassan et al. (2007) and Birch et al. (2007) improved a PBMT system by incorporating syntax in the form of supertags. Marton and Resnik (2008) and Cherry (2008) imposed syntactic constraints by making use of prior linguistic knowledge in the form of syntax analysis. 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.

This paper focuses on incorporating syntactic information into a PBMT model. Our motivation is that PBMT is good at generating translations inherent in the phrase pairs but inefficient at grammatically reordering the target phrases. To deal with this problem, we propose a novel phrasal syntactic category sequence (PSCS) model which allows a PBMT decoder to prefer more grammatical target phrase sequences and better translations.

2 Target Phrase Annotation

In this section, we briefly review the phrase pair extraction algorithm and describe how to assign a syntactic category to each phrase pair.

The basic translation unit of a PBMT model is a phrase pair consisting 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. All pairs of "source word sequence ||| target word sequence" that are consistent with word alignments are collected. Prior to the phrase pair extraction, we use the Berkeley parser[1] (Petrov et al., 2006) to generate the most likely parse tree for each English target sentence in the training corpus.

There are many ways to annotate a phrase pair using a parse tree. Here, we follow the method used in Zollmann and Venugopal (2006). In detail, if the target side of any of these phrase pairs corresponds to a syntactic category of the target side parse tree, we label the phrase pair with that syntactic category. Phrase pairs that do not correspond to a span in the parse tree are given a default category "X". For each phrase pair, we also record the original position of its first and last target word in the target sentence. These position indices will be used in the next section.

For example, given a Chinese-English sentence pair, the English parse tree and the word alignments as shown in Figure 1, we can extract the phrase pairs and the syntactic categories shown in Table 1. Note that if there are any unary rules in the parse tree, we only keep the highest node. So "NP" over the word "this" is kept and "DT" is ignored.

We ran the above procedure on the entire parallel corpus. We may extract the same phrase pair from different parallel sentence pairs whose target side parsing trees are different. So a phrase pair may have multiple syntactic categories. We record all possible syntactic categories and their counts for each phrase pair.

[1] http://code.google.com/p/berkeleyparser/

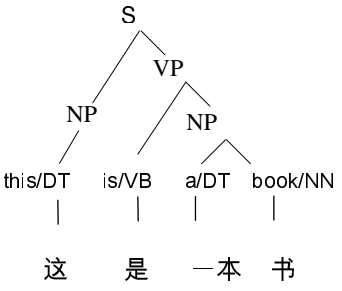

Fig. 1. An example parse tree and word-based alignments

(Zollmann and Venugopal, 2006) used the syntactic category in a hierarchical PBMT model by treating category as non-terminal symbols. In the next section, we propose a novel model which uses the syntactic category in a conventional (i.e., non-hierarchical) PBMT system.

Table 1. Phrase pairs and the syntactic categories extracted from the example in Figure 1

Source phrase	Target phrase	Syntactic category	start	end
这	this	NP	0	0
是	is	VB	1	1
一本	a	DT	2	2
书	book	NN	3	3
这 是	this is	X	0	1
是 一本	is a	X	1	2
一本 书	a book	NP	2	3
这 是 一本	this is a	X	0	2
是 一本 书	is a book	VP	1	3
这 是 一本 书	this is a book	S	0	3

3 PSCS Model

In a PBMT system, the translation candidates are generated from left to right by using a sequence of phrase pairs. Together with the sequence of phrase pairs, a PSCS in the form of $sc_1 sc_2 ,..., sc_n$ is also generated. The variable sc_i stands for the syntactic

category of the *i*-th phrase pair. For example, if we translate a sentence "这 是 一本 书" with the phrase pairs "这 ||| this" and "是 一本 书 ||| is a book" then the corresponding PSCS is "NP VP". By preferring more likely PSCS, one would expect that the output of the decoder will be more grammatical.

We use a bi-gram model to calculate the probability of a PSCS:

$$\begin{aligned} & P(sc_1 sc_2,...,sc_n) \\ &= P(sc_1 | <s>) \cdot P(sc_2 | sc_1) \cdot ,... ,\cdot \\ & \quad P(sc_n | sc_{n-1}) \cdot P(</s> | sc_n) \end{aligned} \tag{1}$$

The start mark <s> and the end mark </s> are used to model how likely sc_1 and sc_n is to occur at the beginning and the end position respectively. The probability is incorporated into the PBMT model log linearly as a new feature.

Our PSCS model is similar to the supertagged language model utilized in Hassan et al. (2007) and Birch et al. (2007). The difference is that they use word-level shallow syntax information in the form of supertags while we use phrase-level full parsing information in the form of syntactic category.

3.1 Dealing with Ambiguity

So far, in this section, we have assumed that each phrase pair used by the decoder has only one syntactic category. However, as we mentioned in section 2, there may be multiple syntactic categories corresponding to one phrase pair.

In order to deal with this ambiguity, one method is to consider each possible syntactic category separately in the decoder. Another is to consider the syntactic category as a hidden variable. In this paper, we use the latter approach simply because it is easy to implement. If the *i*-th phrase pair pp_i has m possible syntactic categories $sc_{i,1}, sc_{i,2}, ..., sc_{i,m}$, and the (*i-1*)-th phrase pair pp_{i-1} has n possible syntactic categories $sc_{i-1,1}, sc_{i-1,2}, ..., sc_{i-1,n}$, then we intuitively replace the log probability $log(P(sc_i|sc_{i-1}))$ in Equation (1) with a linear combination score:

$$\sum_{p=1}^{m} \sum_{q=1}^{n} P(sc_{i,p} \mid pp_i) \cdot P(sc_{i-1,q} \mid pp_{i-1}) \cdot log(P(sc_{i,p}|sc_{i-1,q}))$$

This score is a weighted sum of all possible PSCS bi-gram log probabilities for two contiguous phrase pairs. The weight is empirically set as:

$$P(sc_{i,p} \mid pp_i) \cdot P(sc_{i-1,q} \mid pp_{i-1})$$

3.2 Training of the PSCS Model

Now we describe how to estimate the parameters of the PSCS model. The syntactic category is of phrase-level information, but there is no explicit phrase segmentation in the parallel corpus. This means that we do not have the syntactic category sequence

data that can be directly used to train our PSCS model. Following the phrase extraction method (Koehn, 2007), a heuristic method is used to solve this problem.

We begin with a set of phrase pairs extracted from each sentence pair in the parallel corpus. Each phrase pair is assigned a syntactic category by the method described in section 2. Then we look at the set of phrase pairs, and if any two of them are contiguous in the target side, we extract the syntactic category of these two phrase pairs as a bigram. Figure 2 shows the details of our algorithm. Then we split each bi-gram to get uni-gram samples, which are used to perform data smoothing. Given the collected uni-gram and bi-gram syntactic category samples and their counts, we use the SRI language modeling toolkit[2] to build a bigram PSCS model. The model is smoothed by Witten-Bell discounting.

Input:
  Source sentence *s*
  Target sentences *t*
  Source phrase *sp*
  Target phrase *tp*
Output: PSCS *bigram*
bigram = empty
For each (*t,s*) in the parallel corpus
  For each (sp_i, tp_i) extracted from (*t,s*)
    For each (sp_j, tp_j) extracted from (*t,s*)
    If ($tp_i.end$ +1== $tp_j.start$)
      bigram.add($tp_i.sc$, $tp_j.sc$)
    End
  If($tp_i.start$==0)
  bigram.add(<*s*>, $tp_i.sc$)
  If($tp_i.end$== *t.length-1*
  bigram.add($tp_i.sc$, </*s*>)
  End
End

Fig. 2. Training algorithm for PSCS model

4 Experiments

Our SMT system is based on a fairly typical phrase-based model (Finch and Sumita, 2008). We use a modified training toolkit adapted from the MOSES decoder to train our SMT model. Our decoder can operate on the same principles as the MOSES decoder. The decoder is modified to accommodate our PSCS model. 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). Lexical reordering model is used in our experiments.

[2] http://www-speech.sri.com/projects/srilm/manpages/ngram-count.1.html

Table 2. Corpora statistics

Data	Sentences	Chinese words	English words
Training set	243,698	7,933,133	10,343,140
Development set	1664	38,779	46,387
Test set	1357	32377	42,444
GIGAWORD	19,049,757	-	306,221,306

The translation model was created from the FBIS corpus. We use a 5-gram language model trained with modified Knesser-Ney smoothing. The language model is trained on the target side of the FBIS corpus and the Xinhua news from the GIGAWORD corpus. The development and test sets are from the NIST MT08 evaluation campaign. Table 2 shows the statistics of the corpora used in our experiments.

4.1 Experiments on PSCS Model

As we mentioned in section 2, we use the Berkeley parser to parse the target side of the parallel corpus. Each phrase pair is annotated with the method introduced in section 2. For sentences in which the parser fails to generate a parse tree, we use the default syntactic category X to annotate the phrase pairs. We extracted 21,862,759 phrase pairs in total. There are 14,890,317 phrase pairs whose target side syntactic category is X. The other 6,972,442, or 31%, of the phrase pairs are annotated with linguistic syntactic categories.

Then we build a PSCS model based on the method proposed in section 3. There were 72 kinds of uni-gram syntactic categories and 2478 kinds of bi-gram syntactic categories.

4.2 Experiments on Chinese-English SMT

In order to confirm the effect of our PSCS model, we performed two translation experiments. The first one was a baseline PBMT experiment. In the second experiment, we incorporated our PSCS model into the PBMT system. The evaluation metric is case-sensitive BLEU-4. The results are given in Table 3.

Table 3. Comparison of translation quality

System	**BLEU Score**
PBMT	17.26
PBMT+PSCS	17.92

We were able to achieve an improvement of about 0.7 BLEU point over the baseline PBMT system. This improvement indicates that syntactic categories, even though only 31% of them maintain linguistic meanings, can help select better translation candidates.

5 Conclusion and Future Work

We propose a novel PSCS model to incorporate syntactic information into the conventional PBMT. The PSCS model allows a PBMT decoder to prefer more grammatical target phrase sequences and better translations. Our method is very simple and yields a 0.7 BLEU point improvement when compared to the baseline PBMT system.

We plan to annotate phrase pairs with additional richer syntactic information to obtain further improvements in future work.

Acknowledgment. The work of HIT in this paper is funded by the project of National Natural Science Foundation of China (No. 61173073).

References

Birch, A., Osborne, M., Koehn, P.: CCG Supertags in Factored Translation Models. In: SMT Workshop. ACL (2007)

Cherry, C.: Cohesive phrase-Based decoding for statistical machine translation. In: ACL- HLT (2008)

Chiang, D.: A hierarchical phrase-based model for statistical machine translation. In: ACL (2005)

DeNeefe, S., Knight, K., Wang, W., Marcu, D.: What can syntax-based MT learn from phrase-based MT? In: EMNLP-CoNLL (2007)

Finch, A., Sumita, E.: Dynamic model interpolation for statistical machine translation. In: SMT Workshop (2008)

Galley, M., Graehl, J., Knight, K., Marcu, D., Deneefe, S., Wang, W., Thayer, I.: Scalable inference and training of context-rich syntactic translation models. In: ACL (2006)

Hassan, H., Sima'an, K., Way, A.: Supertagged phrase-based statistical machine translation. In ACL (2007)

Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: HLT-NAACL (2003)

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: ACL demo and poster sessions (2007)

Liu, Y., Liu, Q., Lin, S.: Tree-to-string alignment template for statistical machine translation. In: ACL-COLING (2006)

Liu, Y., Huang, Y., Liu, Q., Lin, S.: Forest-to-string statistical translation rules. In: ACL (2007)

Marcu, D., Wang, W., Echihabi, A., Knight, K.: SPMT: Statistical machine translation with syntactified target language phrases. In: EMNLP (2006)

Marton, Y., Resnik, P.: Soft syntactic constraints for hierarchical phrased-based translation. In: ACL-HLT (2008)

Och, F.: Minimum error rate training in statistical machine translation. In: ACL (2003)

Petrov, S., Barrett, L., Thibaux, R., Klein, D.: Learning accurate, compact, and interpretable tree annotation. In: COLING-ACL (2006)

Quirk, C., Menezes, A., Cherry, C.: Dependency treelet translation: Syntactically informed phrasal SMT. In: ACL (2005)

Xiong, D., Zhang, M., Aw, A., Li, H.: A syntax-driven bracketing model for phrase-based translation. In: ACL-IJCNLP (2009)

Yamada, K., Knight, K.: A syntax-based statistical translation model. In: ACL (2000)

Zhang, M., Jiang, H., Aw, A., Tan, C.L., Li, S.: A tree sequence alignment-based tree-to-tree translation model. In: ACL- HLT (2008)

Zollmann, A., Venugopal, A.: Syntax augmented machine translation via chart parsing. In: SMT Workshop, HLT-NAACL (2006)