

A SP synchronous generation model for statistical machine translation ¹

Jiadong Sun ² , Tiejun Zhao ² and Huashen Liang ²

MOE-MS Key Laboratory of Natural Language Processing and Speech
Harbin Institute of Technology, Harbin, 150001 , China

(Received May 9, 2008, accepted August 6, 2008)

Abstract. We present a Synchronous generation model in the framework of multi-text grammar (MTG). To construct projections between structures at different levels in statistical machine translation (SMT), we give the definitions of Subgraph and Subgraph pairs (SP) in this paper. By the subgraphs of the parse trees, any phrase in the phrase-based models can be generated in this model, syntactic or non-syntactic. To incorporate them into the generation grammar formalism, we propose the operations of addition between graphs called addition. We also give the algorithm of SP extraction. In the experiment, our model outperforms the state-of-the-art Pharaoh by 5.75% and shows better generative ability.

Keywords: structure divergence, structure alignments, complement graph, multi-text grammar.

1. Introduction

Since the Source-Channel model was introduced into machine translation (Brown et al., 1993), Phrase-based models (Marcu and Wong 2002; Koehn et al., 2003; Vogel et al., 2003; Och and Ney, 2004), as an extension of word-based models, exploit the alignment information of the adjacent words as a basic unit, and can produce much better translation results for those consecutive-words units that have been observed in training. But this kind of models cannot robustly perform the reordering task over the phrase level for the syntactical divergent language pairs.

To make the model generalized to unseen or non-syntactic phrases, researchers have presented some syntax-based models (Quirk et al., 2005; Marcu et al., 2006), which induced the syntactic information for the phrases by exploiting tree-to-string rules. But lots of the phrases are not syntactic constituents. To tackle this problem, Marcu et al. (2006) decorated the target language phrases and presented **XRS** rules by introducing non-syntactic symbols to the structure of target parse trees. Liu et al. (2007) presented forest-to-string templates, to describe the alignment information between syntactic source language phrases and target ones. All these approaches produced better translations than state-of-the-art phrasal systems. However, it is hard for these tree-based ordering models to generate and reorder the phrases, which can only be subsumed by the structure across different fragments of the tree structure.

To make the translation model incorporate bilingual syntactic information into statistical translation, and maintain the strengths of the phrase-based models, researchers presented some models, which can be formalized as synchronous generation grammars (Melamed, 2004; Chris Quirk, 2006, Jiadong Sun, et al. 2007) or can be represented in the theory of tree-to-tree transducers (Graehl and Knight, 2004). Ding et al. (2005) induces rules from dependency trees. Compared to the absolute phrasal models, these models improved the translation results, but we noted that isomorphism for the grammar units should be made, but structural divergence between languages has been the major problem for syntax based models.

In order to deal with the problems mentioned above, we propose a novel translation model in this paper.

In section 2, we describe the weighted multi-text grammar (WMTG). To make the model acquire the

¹ Supported by National Natural Science Foundation of China (NSFC:200606010108 and 2006AA01Z150)

² Corresponding author. Tel.: +86-451-86416225-601

E-mail address: jiadongsun@hit.edu.cn.

phrases, we define Sub-graph of the parse trees, and present the tree structure alignment in the units of sub-graphs (Section3).Section 4 addresses the feature function definition for the combination with the WMTG formalism. At last, we describe the experiment details in Section 5. Experiments in the FBI corpus show that our model outperforms the baseline system Pharaoh by 5.75%.

2. The weighted MTG and Structure Divergence

2.1. An introduction for MTG

With sub-graphs (section3), we present a new weighted MTG formalism (multi-tree grammar) in this paper, which can be adapted for any synchronous grammar formalism. This sub-graph based grammar is a generalization of context-free grammar. As follows, we will present a simple introduction of MTG first, in this section. I. Dan Melamed (2003 , 2004), presented a formalism, in which he applied MTG to the statistical machine translation. In this grammar, every production is either a terminal production or a non-terminal production. If there are alignments in, and productions in (1), we know that there are three kinds of languages can be generated synchronously in different lines.

$$\begin{pmatrix} A \\ \varepsilon \\ D \end{pmatrix}; \quad \begin{pmatrix} B \\ C \\ E \end{pmatrix} \tag{1}$$

$$X \Rightarrow AB; Y \Rightarrow C; Z \Rightarrow DED \tag{2}$$

Where we use ε to express there is no symbols to appear in this step of generation. So we will get the following 3 sentences for the different languages : A B; C; D E D; and we can express all of the above in such a way:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \Rightarrow \infty \begin{bmatrix} 1 & 2 \\ 1 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} A & B \\ () & C \\ D & (2) & E \end{bmatrix} \Rightarrow \begin{bmatrix} AB \\ C \\ DED \end{bmatrix} \tag{3}$$

2.2. Structure Divergence

Now we have a look at this symposium in the alignments in the two parse trees and we can't formalize it into any synchronous productions in the MTG, because the nodes {b, c} in Figure1. are generated by

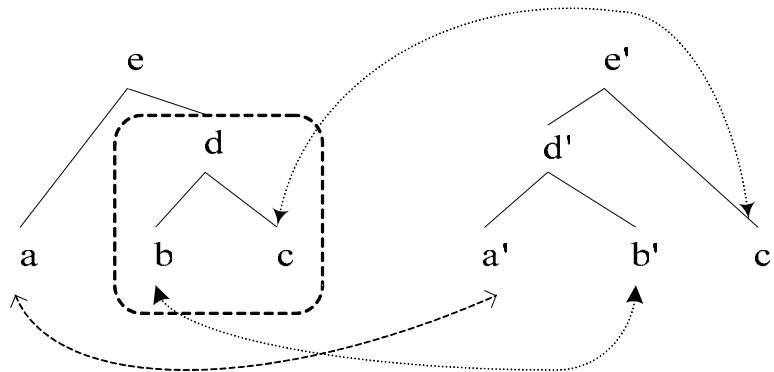


Fig1. The Structure divergence between parse trees

{d}, and the set for the projection of {b, c} is not the projection set of {d}. So in order to get the alignment for structure, we give these definitions as follows in Section 3.

3. Structure Alignments based on sub-graphs

We need a kind of structure, which can construct the alignments of structures and generate all of the phrases in the phrase-based models. We deal with the problem in the way of sub-graphs, not of nodes. On the other hand, suppose that we have the corresponding sub-graphs between source language parse tree and target language parse tree, what kind of methods we need to generate the parse trees of the source language and the target language synchronously? In this section, we will give all of the answers to those questions.

3.1. Sub-graph and Sub-graph Pairs

To make it readable, let's begin with an example for the sub-graph of a parse tree. In Figure 2, a parse tree for an English sentence (ES) is decomposed into two parts of sub-graphs. They are g_1 and g_2 .

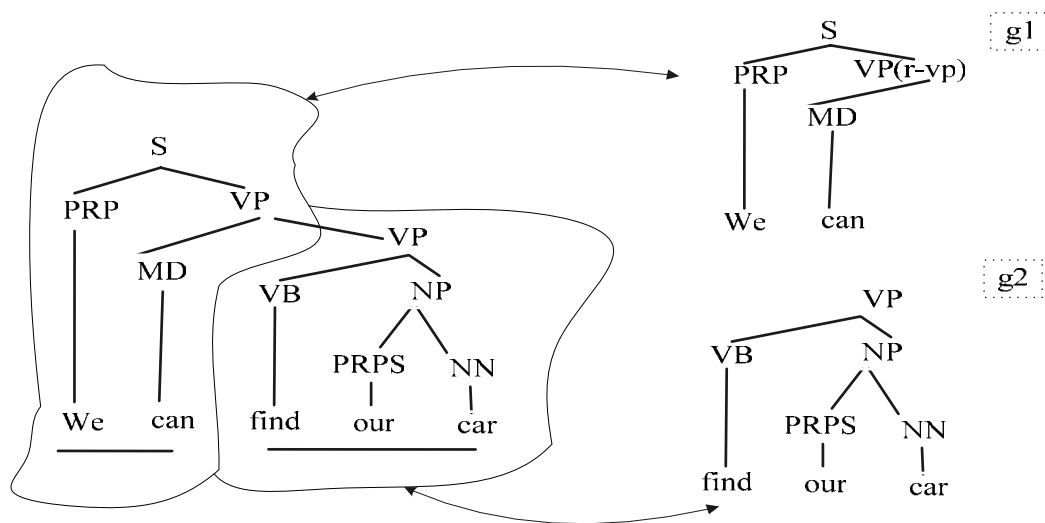


Fig. 2: Decomposition into sub-graphs.

Now, we give the formal definition of subgraph.

Definition 1: Subgraph (S)

A **subgraph** of a parse tree is a triple $\langle g, R, L \rangle$:

1. It is a rooted-tree³ structure;
2. R is the relabeled set, L is the string by the root node.

Sub-graphs in Figure2-b are denoted as follows;

$$S_1 = \langle g_1, \{VP(r-vp)\}, We \ can \rangle \tag{4}$$

$$S_2 = \langle g_2, \{\phi\}, find \ our \ car \rangle \tag{5}$$

Where, $L(g_1) = We \ can$, $R(g_1) = \{Vp(r-vp)\}$, and the index (r-vp) means a sub-graph is detached from this vertex Vp. In Figure2-a, it is noted that the phrase (We can) cannot be captured in the framework presented in Marcu et al., (2006). Here, we decorate the phrase with a sub-graph (g_1).

Definition 2: Sub-graph Pairs(SP)

SP is a vector of three dimensions, $SP = (S, s, \{N, (sq)\})$

1. S is a sub-graph in the source language parse tree, and s is its correspondent part in the target one;
2. N is the number of the symbols in Set L(S) and sq is the proceeding order when the L(S) is mapped according to the order of L(s).

In Figure 3, we give two parse trees, and the alignments between the L (G) and L (g). The whole English sentence is "We can find our car".

³ It is a tree in the theory of graph and may not be a whole fragment of the parse tree.

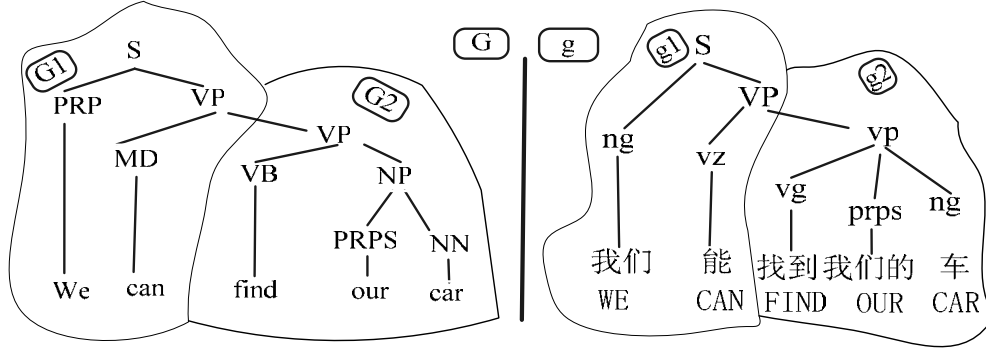


Figure3. Alignments based on sub-graph pairs

Given the tree pair in Figure3, we show two pairs of Sub-graph, which are expressed in the vectors:

$$SP_1 = (S_1, s_1, \{1, (1,2)\});$$

$$SP_2 = (S_2, s_2, \{3, (1,2,3)\}).$$

In fact, If we take the sub-graphs as the basic unit, the alignments for the two parse trees can be acquired at Multi-levels. The sub-graph in the sub-graph pair can be an empty graph.

3.2. Operations of Addition between Sub-graphs

In order to incorporate with the grammar generation formalism, we propose an operation between sub-graphs. We call it Graph Addition. By this operation, a graph can be reconstructed with its sub-graphs.

Definition3: Graph Addition is an operation between sub-graphs.

Given sub-graphs: $S = \langle G, R(G), L(G) \rangle$ and $S' = \langle G', R(G'), L(G') \rangle$, we call S^+ is the sum of them.

If the root of G , $r_G \in L(G')$,

$$S^+ = \langle G \oplus G', \{R(G) \cup R(G')\}, L(G \oplus G') \rangle;$$

If the root of G , $r_G \in R(G')$,

$$S^+ = \langle G \oplus G', \{R(G) \cup R(G') - r_G\}, L(G \oplus G') \rangle.$$

It is noted that the additions just for the vertices in the R or L sets. To make it clear, We give an example for the addition operation: given the S 's in the left part in Figure 3,

$$S^+ = \langle G, \{\phi\}, (we\ can\ find\ the\ car) \rangle$$

3.3. Feature functions for WMSPG

Now, with the structures defined above, we present the weighted SP grammar (WSPG). As we know, the task for the synchronous grammar in SMT is:

$$e = \left\{ \begin{array}{l} \arg \max \{w(D)\} \\ D \text{ s.t. } f(D) = f. \end{array} \right\} \quad (6)$$

To induce weight into the generation process, we define the weight function for the SP and the addition between SPs:

- a weighted SP:

In order to capture the knowledge of appearance for a special SP, we define such feature function: to model the synchronous generation process:

$$f_{1,2}(SP) = \frac{\text{count}(SP)}{\sum_j \text{count}(SP_j)};$$

for f_1 s.t. $S \in SP_j$; and for f_2 $s \in SP_j$

With the $L(S)$ and $L(s)$, we define feature functions:

$$f_3(SP) = \prod_i P(L(S)_i | L(s)_j) \text{ s.t. } L(S)_i \sim L(s)_j$$

$$f_4(SP) = \prod_i P(L(s)_i | L(S)_j) \text{ s.t. } L(S)_i \sim L(s)_j$$

Where the symbol means that the two symbols can be aligned in the SP. The weighted SP is denoted in this way: $SP = (S, s, \{N, (sq)\})\phi$ where

$$\phi(SP) = \frac{\exp[\sum_i \lambda_i f_i(SP)]}{\sum_{j(s_j=s)} \exp[\sum_i \lambda_i f_i(SP_j)]}$$

This final function is given in the framework of log-linear model, in which the parameters λ_i can be trained with the mini-error method (Och and Ney, 2002).

- the weight for Addition between SPs:

In the theory of data generating process, we can have Equation 6 for the combination of the rules in the WMSPG.

$$\begin{bmatrix} S \\ s \\ \sim \end{bmatrix} \xRightarrow{\sum_i R_i} \begin{bmatrix} S^+ \\ s^+ \\ \sim \end{bmatrix} (\phi \prod_{k=i}^j \phi_k) \text{ Such that } \sum_i R_i = \begin{bmatrix} \sum_i S_i \\ \sum_i s_i \\ \sim \end{bmatrix}; \tag{6}$$

In this production, ϕ_i is the weight for SP_i . Consequently, the process of a derivation can also be taken as a process of generation of graphs synchronously.

4. SP Extraction and the Formal SMSPG Framework:

In order to present this SP-based model in the formalism of synchronous grammar, we need another definition. At the same time, we will present the formal SMSPG in this section:

Definition4. **Complement Graph**

Given $SP = SP_1 \oplus SP_2$, we call SP_1 and SP_2 Complement Graph to each other;

For Synchronous generation grammar formalism:

$$\begin{bmatrix} S \\ s \\ \sim \end{bmatrix} \Rightarrow \begin{bmatrix} S_1 & S_2 & \dots & S_n \\ s_1 & s_2 & \dots & s_n \\ \sim_1 & \sim_2 & \dots & \sim_n \end{bmatrix} (\prod_{i=1}^n \phi_i) \tag{7}$$

To acquire all these rules of generations, we can do extraction work. Note that, if a SP is extracted from a parse tree pairs, the remaining graphs can satisfy constrains on SP. So we can do this extraction process on each of them recursively.

5. Experiments

The experiment was conducted for the task of Chinese-to-English translation. We trained the translation model on the FBIS corpus (7.2M+9.2M words).The word alignments are obtained by running GIZA++ (Och and Ney, 2000) on the training set. The parse trees for the source and target languages are obtained with a Chinese parser (H L Cao et al., 2007) and the Collins parser. The Chinese parser’s precision is 85.61%, when tested on the corpus of Penn Chinese Treebank. for the language model, we used the SRI Language Modeling Toolkit to train a trigram model. We used the 2002 NIST MT evaluation test set as our development set, and the 2003 test set as our test set. Our evaluation metric was BLEU (Papineni et al., 2002), as calculated by the NIST script with its default settings, which is to perform case-insensitive

matching of n-grams up to $n = 4$, and to use the shortest (as opposed to nearest) reference sentence for the brevity penalty. The results of the experiments are summarized in Table 1.

Table1: Data characteristics

Corpus	Chinese	English
Train	7.2M	9.2M
Test	23k	25k

Table2: Sample translations

Pharaoh	We can find the book in the room.
SP	We can find the book in the room.
Pharaoh	We can move table in the room.
SP	We can move the table into the room
Pharaoh	With he help, we can move table in room.
SP	With his help, we can move the table into the room.

5.1. The baseline

The system was Pharaoh (Koehn, 2004), which uses a beam search algorithm for decoding. In its model, it takes the following features: language model, phrase translation probability in the two directions, distortion model, word penalty and phrase penalty, all of which can be achieved with the training toolkits distributed by Koehn. The training set and development set mentioned above were used to perform the training task and to tune the feature weights by the minimum error training algorithm. All the other settings were the same as the default ones. SRI Language Model Toolkit was used to train a 3-gram language model. After training, 164 MB language model were obtained.

5.2. Our model

All the common features shared with Pharaoh were trained with the same toolkits and the same corpus. Besides those features, we need train the parameters about sub-graph pairs for our model. The Collins parser and a Chinese parser were used. After processing this corpus, we get a parallel tree corpus. Different numbers of bilingual parse tree pairs were selected to achieve the training set for this part of task (Table 1). To evaluate the result of the translation, the BLEU metric (Papineni et al. 2002) was used.

Table3: System Comparison

System	BLEU
Pharaoh	0.2673 ± 0.0012
SP system	0.2825 ± 0.0016

6. Conclusions

A framework for statistical machine translation is created in this paper. The results of the experiments show that this model gives better performance, compared with the baseline system. This model can incorporate the syntactic information into the process of translation and model the sub-structure projections across the parallel parse trees. The advantage of this frame work lies in that the reordering operations can be

performed at the different levels according to the sub-graph pairs of the bilingual parse trees.

But we should notice that some independent assumptions were made in the decomposition of the parse tree. In the future, a proper method should be introduced into this model to achieve the most possible decomposition of the parse tree. In fact, we can incorporate some other feature functions into the model to generate the multi-texts more effectively.

7. References

- [1] Christoph Tillman. *A projection extension algorithm for statistical machine translation*. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan. June 30-July 4, 2003, 1-8.
- [2] Chris Quirk and Simon Corston-Oliver. The impact of parse quality on syntactically-informed statistical machine translation. In *Proceedings of EMNLP 2006*, Sydney, Australia, July. pages 62–69.
- [3] Daniel Gildea. Loosely tree based alignment for machine translation. In *Proceedings of ACL-03*. 2003
- [4] Daniel Marcu and William Wong. A Phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP'*, Philadelphia, PA. July 6-7, 2002, pages 133-139
- [5] Daniel Marcu, Wei Wang, Abdessamad Echihabi and Kevin Knight. Training tree transducers. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Proceeding (EMNLP)*, Sydney. 2006, pages 44-52.
- [6] David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL 2005*, Ann Arbor, Michigan, June. 2005, pages 263–270.
- [7] Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*. 1997, **23**(3): 3-403.
- [8] Franz J. Och, Hermann Ney. Discriminative training and maximum entropy models. In *Proceedings of ACL 2002*, Hong Kong, October. 2002, pages 440-447.
- [9] Jiadong Sun, Tiejun Zhao and Huashen Liang. Meta-Structure Transformation Model for Statistical Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation of ACL*, 2007, Prague, Czech Republic. pages 64-71.
- [10] Jonathan Graehl Kevin Knight. Training Tree Transducers. In *Proceedings of NAACL-HLT 2004*. pages 105-112.
- [11] Michael John Collins. *Head-driven statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia. 1999.
- [12] I. D. Melamed. Multitext Grammars and SynchronousParsers. *Proceedings of HLT/NAACL*. 2003.
- [13] I. Dan Melamed. Statistical Machine Translation by Parsing. In *Proceedings of the 42nd Annual Conference of the Association for Computational Linguistics (ACL)*, Barcelona, Spain. 2004.
- [14] P. Koehn, Franz Josef Och, Daniel Marcu. Statistical phrase-based translation. *Proceedings of the Conference on Human Language Technology*. Edmonton, Canada. May 27-June 1, 2003, 127-133.
- [15] P. Koehn. Pharaoh: a Beam Search Decoder for Phrase-based Statistical Machine Translation Models. *Meeting of the American Association for machine translation (AMTA)*, Washington DC. Sep./Oct, 2004, pp 115-124.
- [16] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert Merrcer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*. 1993, **19**(2): 263-311.
- [17] Quirk, Chris, Arul Menezes, and Colin Cherry. Dependency Tree Translation. *Microsoft Research Technical Report: MSR-TR-2004-113*.
- [18] Regina Barzilay and Lillian Lee. Learning to paraphrase: An supervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL*. 2003.
- [19] Regina Barzilay and Lillian Lee. Learning to paraphrase: An supervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL*. 2003.
- [20] S.β Nie en, H. Ney: Statistical Machine Translation with Scarce Resources using Morpho-syntactic Information. *Computational Linguistics*. **30**(2): pp. 181-204.

- [21] Yuan Ding and Martha Palmer. Machine translation using probabilistic synchronous dependency insert grammars. In *Proceedings of 43rd Annual Meeting of the NAACL-HLT2004*. pages 273-280.