A Statistical Machine Translation Model Based on a Synthetic Synchronous Grammar

Hongfei Jiang, Muyun Yang, Tiejun Zhao, Sheng Li and Bo Wang
School of Computer Science and Technology
Harbin Institute of Technology
{hfjiang,ymy,tjzhao,lisheng,bowang}@mtlab.hit.edu.cn

Abstract

Recently, various synchronous grammars are proposed for syntax-based machine translation, e.g. synchronous context-free grammar and synchronous tree (sequence) substitution grammar, either purely formal or linguistically motivated. Aiming at combining the strengths of different grammars, we describes a synthetic synchronous grammar (SSG), which tentatively in this paper, integrates a synchronous context-free grammar (SCFG) and a synchronous tree sequence substitution grammar (STSSG) for statistical machine translation. The experimental results on NIST MT05 Chinese-to-English test set show that the SSG based translation system achieves significant improvement over three baseline systems.

1 Introduction

The use of various synchronous grammar based formalisms has been a trend for statistical machine translation (SMT) (Wu, 1997; Eisner, 2003; Galley et al., 2006; Chiang, 2007; Zhang et al., 2008). The grammar formalism determines the intrinsic capacities and computational efficiency of the SMT systems.

To evaluate the capacity of a grammar formalism, two factors, i.e. generative power and expressive power are usually considered (Su and Chang, 1990). The generative power refers to the ability to generate the strings of the language, and the expressive power to the ability to describe the same language with fewer or no extra ambiguities. For the current synchronous grammars based SMT, to some extent, the *generalization ability* of the grammar rules (the usability of the rules for the new sentences) can be considered as a kind of the generative power of the grammar and the *disambiguition ability* to the rule candidates can be considered as an embodiment of expressive power.

However, the *generalization ability* and the *disambiguition ability* often contradict each other in practice such that various grammar formalisms in SMT are actually different trade-off between them. For instance, in our investigations for SMT (Section 3.1), the **F**ormally **SCFG** based hierarchical phrase-based model (hereinafter FSCFG) (Chiang, 2007) has a better generalization capability than a **L**inguistically motivated **STSSG** based model (hereinafter LSTSSG) (Zhang et al., 2008), with 5% rules of the former matched by NIST05 test set while only 3.5% rules of the latter matched by the same test set. However, from expressiveness point of view, the former usually results in more ambiguities than the latter.

To combine the strengths of different synchronous grammars, this paper proposes a statistical machine translation model based on a synthetic synchronous grammar (SSG) which syncretizes FSCFG and LSTSSG. Moreover, it is noteworthy that, from the combination point of view, our proposed scheme can be considered as a novel system combination method which goes beyond the existing post-decoding style combination of N-best hypotheses from different systems.

2 The Translation Model Based on the Synthetic Synchronous Grammar

2.1 The Synthetic Synchronous Grammar

Formally, the proposed Synthetic Synchronous Grammar (SSG) is a tuple

$$G = \langle \Sigma_s, \Sigma_t, N_s, N_t, X, \boldsymbol{P} \rangle$$

where $\Sigma_s(\Sigma_t)$ is the alphabet set of source (target) terminals, namely the vocabulary; $N_s(N_t)$ is the alphabet set of source (target) non-terminals, such

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 125–128,
Suntec, Singapore, 4 August 2009.

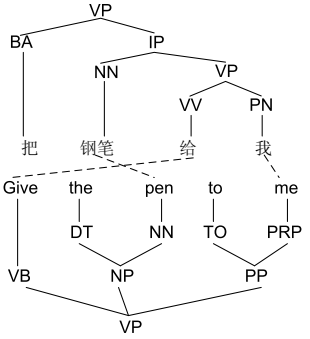

Figure 1: A syntax tree pair example. Dotted lines stands for the word alignments.

as the POS tags and the syntax labels; X represents the special nonterminal label in FSCFG; and P is the grammar rule set which is the core part of a grammar. Every rule r in P is as:

$$r = \langle \alpha, \gamma, A_{NT}, A_T, \bar{\omega} \rangle$$

where $\alpha \in [\{X\}, N_s, \Sigma_s]^+$ is a sequence of one or more source words in Σ_s and nonterminals symbols in $[\{X\}, N_s]$;$\gamma \in [\{X\}, N_t, \Sigma_t]^+$ is a sequence of one or more target words in Σ_t and nonterminals symbols in $[\{X\}, N_t]$; A_T is a many-to-many corresponding set which includes the alignments between the terminal leaf nodes from source and target side, and A_{NT} is a one-to-one corresponding set which includes the synchronizing relations between the non-terminal leaf nodes from source and target side; $\bar{\omega}$ contains feature values associated with each rule.

Through this formalization, we can see that FSCFG rules and LSTSSG rules are both included. However, we should point out that the rules with mixture of X non-terminals and syntactic non-terminals are not included in our current implementation despite that they are legal under the proposed formalism. The rule extraction in current implementation can be considered as a combination of the ones in (Chiang, 2007) and (Zhang et al., 2008). Given the sentence pair in Figure 1, some SSG rules can be extracted as illustrated in Figure 2.

2.2 The SSG-based Translation Model

The translation in our SSG-based translation model can be treated as a SSG derivation. A derivation consists of a sequence of grammar rule applications. To model the derivations as a latent variable, we define the conditional probability distribution over the target translation e and the corresponding derivation d of a given source sentence f as

Input: A source parse tree $T(f_1^J)$
Output: A target translation $\hat{e}$
for $u := 0$ *to* $J-1$ **do**
  for $v := 1$ *to* $J-u$ **do**
    foreach *rule* $r = \langle \alpha, \gamma, A_{NT}, A_T, \bar{\omega} \rangle$ *spanning* $[v, v+u]$ **do**
      if A_{NT} *of* r *is empty* **then**
        Add r into $H[v, v+u]$;
      end
      else
        Substitute the non-terminal leaf node pair (N_{src}, N_{tgt}) with the hypotheses in the hypotheses stack corresponding with N_{src}'s span iteratively.
      end
    end
  end
end
Output the 1-best hypothesis in $H[1, J]$ as the final translation.

Figure 3: The pseudocode for the decoding.

$$(1) \qquad p_\Lambda(\mathbf{d}, \mathbf{e}|\mathbf{f}) = \frac{\exp \sum_k \lambda_k H_k(\mathbf{d}, \mathbf{e}, \mathbf{f})}{\Omega_\Lambda(\mathbf{f})}$$

where H_k is a feature function ,λ_k is the corresponding feature weight and $\Omega_\Lambda(\mathbf{f})$ is a normalization factor for each derivation of $\mathbf{f}$. The main challenge of SSG-based model is how to distinguish and weight the different kinds of derivations . For a simple illustration, using the rules listed in Figure 2, three derivations can be produced for the sentence pair in Figure 1 by the proposed model:

$$d_1 = (R_4, R_1, R_2)$$
$$d_2 = (R_6, R_7, R_8)$$
$$d_3 = (R_4, R_7, R_2)$$

All of them are SSG derivations while d_1 is also a FSCFG derivation, d_2 is also a LSTSSG derivation. Ideally, the model is supposed to be able to weight them differently and to prefer the better derivation, which deserves intensive study. Some sophisticated features can be designed for this issue. For example, some features related with structure richness and grammar consistency[1] of a derivation should be designed to distinguish the derivations involved various heterogeneous rule applications. For the page limit and the fair comparison, we only adopt the conventional features as in (Zhang et al., 2008) in our current implementation.

[1]This relates with reviewers' questions: "can a rule expecting an NN accept an X?" and "... the interaction between the two typed of rules ...". In our study in progress, we would design some features to distinguish the derivation steps which fulfill the expectation or not, to measure how much heterogeneous rules are applied in a derivation and so on.

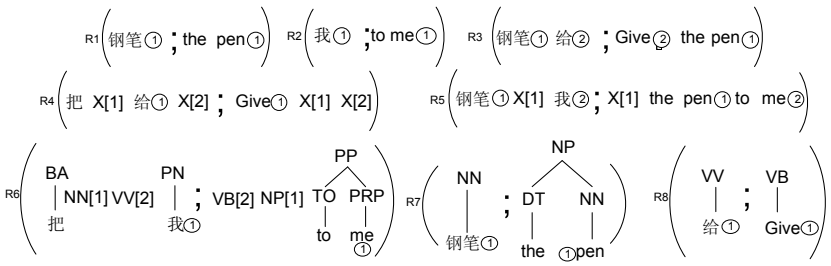

Figure 2: Some synthetic synchronous grammar rules can be extracted from the sentence pair in Figure 1. R_1-R_3 are bilingual phrase rules, R_4-R_5 are FSCFG rules and R_6-R_8 are LSTSSG rules.

2.3 Decoding

For efficiency, our model approximately search for the single 'best' derivation using beam search as

$$(\hat{\mathbf{e}}, \hat{\mathbf{d}}) = \underset{\mathbf{e},\mathbf{d}}{\operatorname{argmax}} \Big\{ \sum_{k} \lambda_k h_k(\mathbf{d}, \mathbf{e}, \mathbf{f}) \Big\}. \quad (2)$$

The major challenge for such a SSG-based decoder is how to apply the heterogeneous rules in a derivation. For example, (Chiang, 2007) adopts a CKY style span-based decoding while (Liu et al., 2006) applies a linguistically syntax node based bottom-up decoding, which are difficult to integrate. Fortunately, our current SSG syncretizes FSCFG and LSTSSG. And the conventional decodings of both FSCFG and LSTSSG are span-based expansion. Thus, it would be a natural way for our SSG-based decoder to conduct a span-based beam search. The search procedure is given by the pseudocode in Figure 3. A hypotheses stack $H[i, j]$ (similar to the "chart cell" in CKY parsing) is arranged for each span $[i, j]$ for storing the translation hypotheses. The hypotheses stacks are ordered such that every span is translated after its possible antecedents: smaller spans before larger spans. For translating each span $[i, j]$, the decoder traverses each usable rule $r = \langle \alpha, \gamma, A_{NT}, A_T, \bar{\omega} \rangle$. If there is no nonterminal leaf node in r, the target side γ will be added into $H[i, j]$ as the candidate hypothesis. Otherwise, the nonterminal leaf nodes in r should be substituted iteratively by the corresponding hypotheses until all nonterminal leaf nodes are processed. The key feature of our decoder is that the derivations are based on synthetic grammar, so that one derivation may consist of applications of heterogeneous rules (Please see d_3 in Section 2.2 as a simple demonstration).

3 Experiments and Discussions

Our system, named HITREE, is implemented in standard C++ and STL. In this section we report on experiments with Chinese-to-English translation base on it. We used FBIS Chinese-to-English parallel corpora (7.2M+9.2M words) as the training data. We also used SRI Language Modeling Toolkit to train a 4-gram language model on the Xinhua portion of the English Gigaword corpus(181M words). NIST MT2002 test set is used as the development set. The NIST MT2005 test set is used as the test set. The evaluation metric is case-sensitive BLEU4. For significant test, we used Zhang's implementation (Zhang et al., 2004)(confidence level of 95%). For comparisons, we used the following three baseline systems:

LSTSSG An in-house implementation of linguistically motivated STSSG based model similar to (Zhang et al., 2008).

FSCFG An in-house implementation of purely formally SCFG based model similar to (Chiang, 2007).

MBR We use an in-house combination system which is an implementation of a classic sentence level combination method based on the Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004).

	Extracted(k)	**S**cored(k)(**S**/**E**%)	**F**iltered(k)(**F**/**S**%)
BP	11,137	4,613(41.4%)	323(0.5%)
LSTSSG	45,580	28,497(62.5%)	984(3.5%)
FSCFG	59,339	25,520(43.0%)	1,266(5.0%)
HITREE	93,782	49,404(52.7%)	1,927(3.9%)

Table 1: The statistics of the counts of the rules in different phases. 'k' means one thousand.

3.1 Statistics of Rule Numbers in Different Phases

Table 1 summarizes the statistics of the rules for different models in three phases: after extraction (*Extracted*), after scoring(*Scored*), and after filtering (*Filtered*) (filtered by NIST05 test set just, similar to the filtering step in phrase-based SMT system). In *Extracted* phase, FSCFG

ID	System	BLEU4	#of used rules(k)
1	LSTSSG	0.2659±0.0043	984
2	FSCFG	0.2613±0.0045	1,266
3	HITREE	0.2730±0.0045	1,927
4	MBR(1,2)	0.2685±0.0044	–

Table 2: The Comparison of LSTSSG, FSCFG ,HITREE and the MBR.

has obvious more rules than LSTSSG. However, in *Scored* phase, this situation reverses. Interestingly, the situation reverses again in *Filtered* phase. The reasons for these phenomenons are that FSCFG abstract rules involves high-degree generalization. Each FSCFG abstract rule averagely have several duplicates[2] in the extracted rule set. Then, the duplicates will be discarded during scoring. However, due to the high-degree generalization , the FSCFG abstract rules are more likely to be matched by the test sentences. Contrastively, LSTSSG rules have more diversified structures and thus weaker generalization capability than FSCFG rules. From the ratios of two transition states, Table 1 indicates that HITREE can be considered as compromise of FSCFG between LSTSSG.

3.2 Overall Performances

The performance comparison results are presented in Table 2. The experimental results show that the SSG-based model (HITREE) achieves significant improvements over the models based on the two isolated grammars: FSCFG and LSTSSG (both $p < 0.001$). From combination point of view, the newly proposed model can be considered as a novel method going beyond the conventional post-decoding style combination methods. The baseline Minimum Bayes Risk combination of LSTSSG based model and FSCFG based model ($MBR(1,2)$) obtains significant improvements over both candidate models (both $p < 0.001$). Meanwhile, the experimental results show that the proposed model outperforms $MBR(1,2)$ significantly ($p < 0.001$). These preliminary results indicate that the proposed SSG-based model is rather promising and it may serve as an alternative, if not superior, to current combination methods.

4 Conclusions

To combine the strengths of different grammars, this paper proposes a statistical machine translation model based on a synthetic synchronous grammar (SSG) which syncretizes a purely formal synchronous context-free grammar (FSCFG) and a linguistically motivated synchronous tree sequence substitution grammar (LSTSSG). Experimental results show that SSG-based model achieves significant improvements over the FSCFG-based model and LSTSSG-based model.

In the future work, we would like to verify the effectiveness of the proposed model on various datasets and to design more sophisticated features. Furthermore, the integrations of more different kinds of synchronous grammars for statistical machine translation will be investigated.

[2]Rules with identical source side and target side are duplicated.

Acknowledgments

This work is supported by the Key Program of National Natural Science Foundation of China (60736014), and the Key Project of the National High Technology Research and Development Program of China (2006AA010108).

References

David Chiang. 2007. Hierarchical phrase-based translation. In *computational linguistics*, 33(2).

Jason Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *Proceedings of ACL 2003*.

Galley, M. and Graehl, J. and Knight, K. and Marcu, D. and DeNeefe, S. and Wang, W. and Thayer, I. 2006. Scalable inference and training of context-rich syntactic translation models In *Proceedings of ACL-COLING*.

S. Kumar and W. Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *HLT-04*.

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

Keh-Yin Su and Jing-Shin Chang. 1990. Some key Issues in Designing Machine Translation Systems. *Machine Translation*, 5(4):265-300.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377-403.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of LREC 2004*, pages 2051-2054.

Min Zhang, Hongfei Jiang, Ai Ti AW, Haizhou Li, Chew Lim Tan and Sheng Li. 2008. A tree sequence alignment-based tree-to-tree translation model. In *Proceedings of ACL-HLT*.