Modeling Syntactic and Semantic Structures in Hierarchical Phrase-based Translation

Junhui Li
University of Maryland
College Park, USA
lijunhui@umiacs.umd.edu

Philip Resnik
University of Maryland
College Park, USA
resnik@umd.edu

Hal Daumé III
University of Maryland
College Park, USA
hal@umiacs.umd.edu

Abstract

Incorporating semantic structure into a linguistics-free translation model is challenging, since semantic structures are closely tied to syntax. In this paper, we propose a two-level approach to exploiting predicate-argument structure reordering in a hierarchical phrase-based translation model. First, we introduce linguistically motivated constraints into a hierarchical model, guiding translation phrase choices in favor of those that respect syntactic boundaries. Second, based on such translation phrases, we propose a predicate-argument structure reordering model that predicts reordering not only between an argument and its predicate, but also between two arguments. Experiments on Chinese-to-English translation demonstrate that both advances significantly improve translation accuracy.

1 Introduction

Hierarchical phrase-based (HPB) translation models (Chiang, 2005; Chiang, 2007) that utilize synchronous context free grammars (SCFG) have been widely adopted in statistical machine translation (SMT). Although formally syntactic, such models rarely respect linguistically-motivated syntax, and have no formal notion of semantics. As a result, they tend to produce translations containing both grammatical errors and semantic role confusions. Our goal is to take advantage of syntactic and semantic parsing to improve translation quality of HPB translation models. Rather than introducing semantic structure into the HPB model directly, we construct an improved translation model by incorporating linguistically motivated *syntactic constraints* into a standard HPB model. Once the translation phrases are linguistically constrained, we are able to propose a *predicate-argument reordering model*. This reordering model aims to solve two problems: ensure that arguments are ordered properly after translation, and to ensure that the proper argument structures even *exist,* for instance in the case of PRO-drop languages. Experimental results on Chinese-to-English translation show that both the hard syntactic constraints and the predicate-argument reordering model obtain significant improvements over the syntactically and semantically uninformed baseline.

In principle, semantic frames (or, more specifically, predicate-argument structures: PAS) seem to be a promising avenue for translational modeling. While languages might diverge syntactically, they are less likely to diverge semantically. This has previously been recognized by Fung et al. (2006), who report that approximately 84% of semantic role mappings remained consistent across translations between English and Chinese. Subsequently, Zhuang and Zong (2010) took advantage of this consistency to jointly model semantic frames on Chinese/English bitexts, yielding improved frame recognition accuracy on both languages.

While there has been some encouraging work on integrating syntactic knowledge into Chiang's HPB model, modeling semantic structure in a linguistically naive translation model is a challenge, because the semantic structures themselves are syntactically motivated. In previous work, Liu and Gildea (2010) model the reordering/deletion of source-side semantic roles in a tree-to-string translation model. While it is natural to include semantic structures in a tree-based translation model, the effect of semantic structures is presumably limited, since tree templates themselves have already encoded semantics to some

Proceedings of NAACL-HLT 2013, pages 540–549,
Atlanta, Georgia, 9–14 June 2013. ©2013 Association for Computational Linguistics

extent. For example, template *(VP (VBG giving) NP#1 NP#2)* entails *NP#1* as *receiver* and *NP#2* as *thing given*. Xiong et al. (2012) model the reordering between predicates and their arguments by assuming arguments are translated as a unit. However, they only considered the reordering between arguments and their predicates.

2 Syntactic Constraints for HPB Translation Model

In this section, we briefly review the HPB model, then present our approach to incorporating syntactic constraints into it.

2.1 HPB Translation Model

In HPB models, synchronous rules take the form $X \rightarrow \langle \gamma, \alpha, \sim \rangle$, where X is the non-terminal symbol, γ and α are strings of lexical items and non-terminals in the source and target side, respectively, and $\sim$ indicates the one-to-one correspondence between non-terminals in γ and α. Each such rule is associated with a set of translation model features $\{\phi_i\}$, including phrase translation probability $p\left(\alpha \mid \gamma\right)$ and its inverse $p\left(\gamma \mid \alpha\right)$, the lexical translation probability $p_{lex}\left(\alpha \mid \gamma\right)$ and its inverse $p_{lex}\left(\gamma \mid \alpha\right)$, and a rule penalty that affects preference for longer or shorter derivations. Two other widely used features are a target language model feature and a target word penalty.

Given a derivation d, its translation probability is estimated as:

$$P\left(d\right) \propto \prod_i \phi_i \left(d\right)^{\lambda_i} \qquad (1)$$

where λ_i is the corresponding weight of feature ϕ_i. See (Chiang, 2007) for more details.

2.2 Syntactic Constraints

Translation rules in an HPB model are extracted from *initial phrase* pairs, which must include at least one word inside one phrase aligned to a word inside the other, such that no word inside one phrase can be aligned to a word outside the other phrase. It is not surprising to observe that initial phrases frequently are non-intuitive and inconsistent with linguistic constituents, because they are based only on statistical word alignments. Nothing in the framework actually requires linguistic knowledge.

Koehn et al. (2003) conjectured that such non-intuitive phrases do not help in translation. They tested this conjecture by restricting phrases to syntactically motivated constituents on both the source and target side: only those initial phrase pairs are subtrees in the derivations produced by the model. However, their phrase-based translation experiments (on Europarl data) showed the restriction to syntactic constituents is actually harmful, because too many phrases are eliminated. The idea of hard syntactic constraints then seems essentially to have been abandoned: it doesn't appear in later work.

On the face of it, there are many possible reasons Koehn et al. (2003)'s hard constraints did not work, including, for example, tight restrictions that unavoidably exclude useful phrases, and practical issues like the quality of parse trees. Although ensuing work moved in the direction of *soft* syntactic constraints (see Section 6), our ultimate goal of capturing predicate-argument structure requires linguistically valid syntactic constituents, and therefore we revisit the idea of hard constraints, avoiding problems with their strictness by relaxing them in three ways.

First, requiring source phrases to be subtrees in a linguistically informed syntactic parse eliminates many reasonable phrases. Consider the English-Chinese phrase pair ⟨*the red car, hongse de qiche*⟩.[1] It is easily to get a translation entry for the whole phrase pair. By contrast, the phrase pair ⟨*the red, hongse de*⟩ is typically excluded because it does not correspond to a complete subtree on the source side. Yet translating *the red* is likely to be more useful than translating *the red car*, since it is more general: it can be followed by any other noun translation. To this end, we relax the syntactic constraints by allowing phrases on the source side corresponding to either *one subtree* or *sibling subtrees* with a common parent node in the syntactic parse. For example, *the red* in Figure 1(a) is allowed since it spans two subtrees that have a common parent node *NP*.

Second, we might still exclude useful phrases because the syntactic parses of some languages, like Chinese, prefer deep trees, resulting in a head and its modifiers being distributed across multiple structural levels. Consider the English sentence *I still*

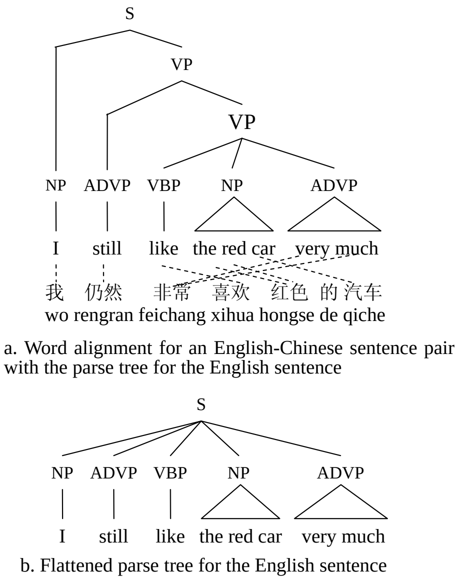
[1] We use English as source language for better readability.

like the red car very much and its syntactic structure as shown in Figure 1(a). Phrases *I still*, *still like*, *I still like* are not allowed, since they don't map to either a subtree or sibling subtrees. Logically, however, it might make sense not just to include phrases mapping to (sibling) subtrees, but to include phrases mapping to subtrees with the same head. To this end, we flatten the syntactic parse so that a head and all its modifiers appear at the same level. Another advantage of this flattened structure is that flattened trees are more reliable than unflattened ones, in the sense that some bracketing errors in unflattened trees can be eliminated during tree flattening. Figure 1(b) illustrates flattening a syntactic parse by moving the head (*like*) and all its modifiers (*I*, *still*, *the red car*, and *very much*) to the same level.

Third, initial phrase pair extraction in Chiang's HPB generates a very large number of rules, which makes training and decoding very slow. To avoid this, a widely used strategy is to limit initial phrases to a reasonable length on either side during rule extraction (e.g., 10 in Chiang (2007)). A corresponding constraint to speed up decoding prohibits any *X* from spanning a substring longer than a fixed length, often the same as the maximum phrase length in rule extraction. Although the initial phrase length limitation mainly keeps non-intuitive phrases out, it also closes the door on some useful phrases. For example, a translation rule ⟨*I still like X, wo rengran xihuan X*⟩ will be prohibited if the non-terminal *X* covers 8 or more words. In contrast, our hard constraints have already filtered out dominating non-intuitive phrases; thus there is more room to include additional useful phrases. As a result, we can switch off the constraints on initial phrase length in both training and decoding.

2.3 Reorderable Glue Rules

In decoding, if no good rule (e.g., a rule whose left-hand side is *X*) can be applied or the length of the potential source span is larger than a pre-defined length, a glue rule (either $S \rightarrow \langle X_1, X_1 \rangle$ or $S \rightarrow \langle S_1 X_2, S_1 X_2 \rangle$) will be used to simply stitch two consequent translated phrases together in monotonic way. This will obviously prevent some reasonable translation derivations because in certain cases, the order of phrases may be inverted on the target side. Moreover, even that the syntactic constraints discussed above make translation node Xs are syntactically informed, stitching translated phrases from left to right will unavoidably generate non-syntactically informed node Ss. For example, the combination of $X\ (like)$ and $X\ (the)$ does not make much sense in linguistic perspective.



a. Word alignment for an English-Chinese sentence pair with the parse tree for the English sentence

b. Flattened parse tree for the English sentence

Figure 1: Example of flattening parse tree.

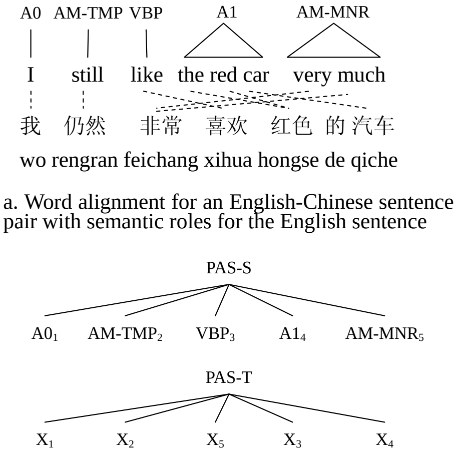
Alternatively, we replace glue rules of HPB with reorderable ones:

- $T \rightarrow \langle X_1, X_1 \rangle$
- $T \rightarrow \langle T_1 T_2, T_1 T_2 \rangle$
- $T \rightarrow \langle T_1 T_2, T_2 T_1 \rangle$

where the second (third) rule combines two translated phrases in a monotonic (inverted) way. Specifically, we set the translation probability of the first translation rule as 1 while estimating the probabilities of the other two rules from training data. In both training and decoding, we require the phrases covered by T to satisfy our syntactic constraints. Therefore, all translation nodes (both Xs and Ts) in derivations are syntactically informed, providing room to explore PAS reordering in HPB model.

3 PAS Reordering Model

Ideally, we aim to model PAS reordering based on the true semantic roles of both the source and target side, as to better cater not only consistence but



a. Word alignment for an English-Chinese sentence pair with semantic roles for the English sentence

b. PAS-S and PAS-T for predicate *like*

Figure 2: Example of PAS on both the source and target side. Items are aligned by indices.

divergence between semantic frames of the source and target language. However, considering there is no efficient way of jointly performing MT and SRL, accurate SRL on target side can only be done after translation. Similar to related work (Liu and Gildea, 2010; Xiong et al., 2012), we obtain the PAS of the source language (PAS-S) via a shallow semantic parser and project the PAS of the target language (PAS-T) using the word alignment derived from the translation process. Specifically, we use PropBank standard (Palmer et al., 2005; Xue, 2008) which defines a set of numbered core arguments (i.e., A0-A5) and adjunct-like arguments (e.g., AM-TMP for temporal, AM-MNR for manner). Figure 2(b) shows an example of PAS projection from source language to target language.[2] The PAS reordering model describes the probability of reordering *PAS-S* into *PAS-T*. Given a predicate p, it takes the following form:

$$P\left(\text{PAS-T} \mid \text{PAS-S, PRE=}p\right) \tag{2}$$

Note that cases for *untranslated roles* can be naturally reflected in our PAS reordering model. For example, if the argument I_{A0} is untranslated in Figure 2, its PAS-T will be $X_2X_5X_3X_4$.

3.1 Probability Estimation

While it is hard and unnecessary to translate a predicate and all its associated arguments with one rule, especially if the sentence is long, a practicable way, as most decoders do, is to translate them in multiple level rules. In addition, some adjunct-like arguments are optional, or structurally dispensable part of a sentence, which may result in data sparsity issue. Based on these observations, we decompose Formula 2 into two parts: *predicate-argument reordering* and *argument-argument reordering*.

Predicate-Argument Reordering estimates the reordering probability between a predicate and one of its arguments. Taking predicate *like* and its argument A1 *the red car* in Figure 2(a) as an example, the predicate-argument pattern on the source side (PA-S) is *VBP*$_1$ *A1*$_2$ while the predicate-argument pattern on the target side (PA-T) is X_1X_2. The reordering probability is estimated as:

$$P_{\text{P-A}}\left(\text{PA-T=X}_1\text{ X}_2 \mid \text{PA-S=VBP}_1\text{ A1}_2\text{, PRE=like}\right) = \frac{Count\left(\text{PA-T=X}_1\text{ X}_2\text{, PA-S=VBP}_1\text{ A1}_2\text{, PRE=like}\right)}{\sum_{\mathcal{T}\in\Phi(\text{PA-S})} Count\left(\text{PA-T=}\mathcal{T}\text{, PA-S=VBP}_1\text{ A1}_2\text{, PRE=like}\right)} \tag{3}$$

where $\Phi\left(PA\text{-}S\right)$ enumerates all possible reorderings on the target side. Moreover, we take the predicate lexicon of predicate into account. To avoid data sparsity, we set a threshold (e.g., 100) to retain frequent predicates. For infrequent predicates, their probabilities are smoothed by replacing predicate lexicon with its POS. Finally, if source side patterns are infrequent (e.g., less than 10) for frequent predicates, their probabilities are smoothed as well with the same way.

Argument-Argument Reordering estimates the reordering probability between two arguments, i.e., argument-argument pattern on the source side (AA-S) and its counterpart on the target side (AA-T). However, due to that arguments are driven and pivoted by their predicates, we also include predicate in patterns of AA-S and AA-T. Let's revisit Figure 2(a). A1 *the red car* and AM-MNR *very much* are inverted on the target side, whose probability is estimated as:

$$P_{\text{A-A}}\left(\text{AA-T=X}_3\text{ X}_1\text{ X}_2 \mid \text{AA-S=VBP}_1\text{ A1}_2\text{ AM-MNR}_3\text{, PRE=like}\right) \tag{4}$$

Similarly we smooth the probabilities by distinguishing frequent predicates from infrequent ones,

[2] In PAS-S, we use parts-of-speech (POS) of predicates to distinguish different types of verbs since the semantic structures of Chinese adjective verbs are different from those of others.

as well as frequent patterns from infrequent ones.

3.2 Integrating the PAS Reordering Model into the HPB Model

We integrate the PAS reordering model into the HPB SMT by adding a new feature into the log-linear translation model. Unlike the conventional phrase and lexical translation features whose values are phrase pair-determined and thus can be calculated offline, the value of the PAS reordering model can only be obtained with being aware of the predicate-argument structures a hypothesis may cover. Before we present the algorithm of integrating the PAS reordering model, we define a few functions by assuming p for a predicate, a for an argument, and H for a hypothesis:

- $\mathcal{A}(i, j, p)$: returns arguments of p which are fully located within the span from word i to j on the source side. For example, in Figure 2, $\mathcal{A}(4, 8, like) = \{A1, AM\text{-}MRN\}$.[3]
- $\mathcal{B}(i, j, p)$: returns *true* if p is located within $[i, j]$; otherwise returns *false*.
- $\mathcal{C}(a, p)$: returns *true* if predicate-argument reordering for a and p has not calculated yet; otherwise returns *false*.
- $\mathcal{D}(a_1, a_2, p)$: returns *true* if argument-argument reordering for p's arguments a_1 and a_2 has not calculated yet; otherwise returns *false*.
- $\mathcal{P}_{P\text{-}A}(H, a, p)$: according to Eq. 3, returns the probability of predicate-argument reordering of a and p, given a and p are covered by H. The positional relation of a and p on the target side can be detected according to translation derivation of H.
- $\mathcal{P}_{A\text{-}A}(H, a_1, a_2, p)$: according to Eq. 4, returns the probability of argument-argument reordering of p's arguments a_1 and a_2, given a_1, a_2 and p are covered by H.

Algorithm 1 integrates the PAS reordering model into a CKY-style decoder whenever a new hypothesis is generated. Given a hypothesis H, it first looks for predicates and their arguments which are covered by H (line 2-3). Then it respectively calculates the probabilities of predicate-argument reordering and argument-argument reordering(line 4-7).

Algorithm 1: Integrating the PAS reordering model into a CKY-style decoder

```
Input: Sentence $f$ in the source language
       Predicate-Argument Structures of $f$
       Hypothesis $H$ spanning from word $i$ to $j$
Output: Log-Probability of the PAS reordering model
1. set $prob$ = 0.0
2. for predicate $p$ in $f$, such that $\mathcal{B}(i, j, p)$ is $true$
3.     $ARG = \mathcal{A}(i, j, p)$
4.     for $a \in ARG$ such that $\mathcal{C}(a, p)$ is $true$
5.         $prob$ += $\log \mathcal{P}_{P\text{-}A}(H, a, p)$
6.     for $a_1, a_2 \in ARG$ such that $a_1 \neq a_2$ and
                                   $\mathcal{D}(a_1, a_2, p)$ is $true$
7.         $prob$ += $\log \mathcal{P}_{A\text{-}A}(H, a_1, a_2, p)$
8. return $prob$
```

4 Experiments

We have presented our two-level approach to incorporating syntactic and semantic structures in a HPB system. In this section, we test the effect of such structural information on a Chinese-to-English translation task. The baseline system is a reproduction of Chiang's (2007) HPB system. The bilingual training data contains 1.5M sentence pairs with 39.4M Chinese words and 46.6M English words.[4] We obtain the word alignments by running GIZA++ (Och and Ney, 2000) on the corpus in both directions and applying "grow-diag-final-and" refinement (Koehn et al., 2003). We use the SRI language modeling toolkit to train a 5-gram language model on the Xinhua portion of the Gigaword corpus and standard MERT (Och, 2003) to tune the feature weights on the development data.

To obtain syntactic parse trees for instantiating syntactic constraints and predicate-argument structures for integrating the PAS reordering model, we first parse the source sentences with the Berkeley Parser (Petrov and Klein, 2007) trained on Chinese TreeBank 6.0 and then ran the Chinese semantic role

[3] The hard constraints make sure a valid source text span would never fully cover some roles while partially cover other roles. For example, phrases *like the red*, *the read car very* in Figure 1 are invalid.

[4] This dataset includes LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06

	System	MT 02	MT 04	MT 05	Ave.
	base HPB	40.00	35.33	32.97	36.10
max-phrase-length=10 max-char-span=10	+ basic constraints + unflattened tree	33.90	32.00	29.83	31.91
	+ our constraints + unflattened tree	38.47	34.51	32.15	35.04
	+ our constraints + flattened tree	38.55	35.38	32.44	35.46
max-phrase-length=∞ max-char-span=∞	+ basic constraints + unflattened tree	35.38	32.89	30.42	32.90
	+ our constraints + unflattened tree	39.41	36.02	33.21	36.21
	+ our constraints + flattened tree	**40.01**	**36.24**	**33.65**	**36.71**

Table 1: Effects of hard constraints. Here max-phrase-length is for maximum initial phrase length in training and max-char-span for maximum phrase length can be covered by non-terminal *X* in decoding.

labeler (Li et al., 2010) on all source parse trees to annotate semantic roles for all verbal predicates.

We use the 2003 NIST MT evaluation test data (919 sentence pairs) as the development data, and the 2002, 2004 and 2005 NIST MT evaluation test data (878, 1788 and 1082 sentence pairs, respectively) as the test data. For evaluation, the NIST BLEU script (version 11b) is used to calculate the NIST BLEU scores, which measures case-insensitive matching of n-grams with n up to 4. To test whether a performance difference is statistically significant, we conduct significance tests following the paired bootstrapping approach (Koehn, 2004).

4.1 Effects of Syntactic Constraints

We have also tested syntactic constraints that simply require phrases on the source side to map to a subtree (called basic constraints). Similar to requiring initial phrases on the source side to satisfy the constraints in training process, we only perform chart parsing on text spans which satisfy the constraints in decoding process. Table 1 shows the results of applying syntactic constraints with different experimental settings. From the table, we have the following observations.

- Consistent with the conclusion in Koehn et al. (2003), using the basic constraints is harmful to HPB. Fortunately, our constraints consistently work better than the basic constraints.
- Relaxing maximum phrase length in training and maximum char span length in decoding, we obtain an average improvement of about 1.0∼1.2 BLEU points for systems with both basic constraints and our constraints. It is worth noting that after relaxing the lengths, the system with our constraints performs on a par with the base HPB system (e.g., 36.21 *vs.* 36.10).
- Flattening parse trees further improves 0.4∼0.5 BLEU points on average for systems with our syntactic constraints. Our final system with constraints outperforms the base HPB system with an average of 0.6 BLEU points improvement (36.71 *vs.* 36.10).

System	MT 02	MT 04	MT 05	Ave.
base HPB	40.00	35.33	32.97	36.10
+our constraints	40.01	36.24^{++}	33.65^{+}	36.71
with reorderable glue rules	$\mathbf{40.70}^{+}$	36.00^{+}	33.67^{+}	36.79
+PAS model	40.41^{+}	$\mathbf{36.73}^{++**}$	$\mathbf{34.24}^{++*}$	**37.13**

Table 2: Effects of reorderable glue rules and the PAS reordering model. +/++: significant over base HPB at 0.05/0.01; */**: significant over the system with reorderable glue rules at 0.05/0.01.

Another advantage of applying syntactic constraints is efficiency. By comparing the base HPB system and the system with our syntactic constraints (i.e., the last row in Table 1), it is not surprising to observe that the size of rules extracted from training data drops sharply from 193M in base HPB system to 60M in the other. Moreover, the system with constraints needs less decoding time than base HPB does. Observation on 2002 NIST MT test data (26 words per sentence on average) shows that basic HPB system needs to fill 239 cells per sentence on average in chart parsing while the other only needs to fill 108 cells.

4.2 Effects of Reorderable Glue Rules

Based on the system with our syntactic constraints and relaxed phrase lengths in training and decoding, we replace traditional glue rules with reorderable glue rules. Table 2 shows the results, from which we find that the effect of reorderable glue rules is elusive: surprisingly, it achieves 0.7 BLEU points

sentence length	1-10	11-20	21-30	31-40	41+	all
sentence count	337	1001	1052	768	590	3748
base HPB	**32.21**	37.51	36.71	34.96	35.00	35.73
+our constraints	31.70	**37.57**	**37.10**	**36.20**$^{++}$	**35.78**$^{++}$	**36.39**$^{++}$

Table 3: Experimental results over different sentence length on the three test sets. +/++: significant over base HPB at 0.05/0.01.

improvement on NIST MT 2002 test set while having negligible or even slightly negative impact on the other two test sets. The reason of reorderable glue rules having limited influence on translation results over monotonic only glue rules may be due to that the monotonic reordering overwhelms the inverted one: estimated from training data, the probability of the monotonic glue rule is 95.5%.

4.3 Effects of the PAS Reordering Model

Based on the system with reorderable glue rules, we examine whether the PAS reordering model is capable of improving translation performance. The last row in Table 2 presents the results . It shows the system with the PAS reordering model obtains an average of 0.34 BLEU points over the system without it (e.g., 37.13 *vs.* 36.79). It is interesting to note that it achieves significant improvement on NIST MT 2004 and 2005 test sets ($p < 0.05$) while slightly lowering performance on NIST MT 2002 test set ($p > 0.05$): the surprising improvement of applying reorderable glue rules on NIST MT 2002 test set leaves less room for further improvement. Finally, it shows we obtain an average improvement of 1.03 BLEU points on the three test sets over the base HPB system.

5 Discussion and Future Work

The results in Table 1 demonstrate that significant and sometimes substantial gains over baseline can be obtained by incorporating hard syntactic constraints into the HPB model. Due to the capability of translation phrases of arbitrary length, we conjecture that the improvement of our system over the baseline HPB system mostly comes from long sentences. To test the conjecture, we combine all test sentences in the three test sets and group them in terms of sentence length. Table 3 presents the sentence distribution and BLEU scores over different length. The results validate our assumption that the system with constraints outperforms the base systems on long sentences (e.g., sentences with 20+ words).

Figure 3 displays a translation example which shows the difference between the base HPB system and the system with constraints. The inappropriate translation of the base HPB system can be mainly blamed on the rule $\langle X_{[2,5]} \rightarrow \text{的}_2\,\text{发展}_3\,X_{[4,5]}，X_{[4,5]}\,the\,development\,of \rangle$, where $\text{的}_2\ \text{发展}_3$, a part of the subtree $[0,3]$ spanning from word 0 to 3, is translated immediately to the right of $X_{[2,5]}$, making a direct impact that subtree $[0,3]$ is translated discontinuously on the target side. On the contrary, we can see that our constraints are able to help select appropriate phrase segments with respect to its syntactic structure.

Although our syntactic constraints apply on the source side, they are completely ignorant of syntax on the target side, which might result in excluding some useful translation rules. Let's revisit the sentence in Figure 3, where we can see that a transition rule spanning from word 0 to 5, say $\langle X_{[0,5]} \rightarrow X_{[0,3]}\,\text{是}_4\,\text{取决于}_5，X_{[0,3]}\,depends\,on \rangle$ is intuitive: the syntactic structure on the target side satisfies the constraints, although that of the source side doesn't. One natural extension of this work, therefore, would be to relax the constraints by including translation rules whose syntactic structure of either the source side or the target side satisfies the constraints.

To illustrate how the PAS reordering model impacts translation output, Figure 4 displays two translation examples of systems with or without it. The predicate 传递*/convey* in the first example has three core arguments, i.e., A0, A2, and A1. The difference between the two outputs is the reordering of A1 and A2 while the PAS reordering model gives priority to pattern *VV A1 A2*. In the second example, we clearly observe two serious translation errors in the system without PAS reordering model: 他们/them$_{A1}$ is untranslated; 中国/china$_{A0}$ is moved to the immediate right of predicate 允许*/allow* and plays as direct object.

Including the PAS reordering model improves the BLEU scores. One further direction to refine the approach is to alleviate verb sparsity via verb classes. Another direction is to include useful context in estimating reordering probability. For example, the content of a temporal argument AM-TMP can be a

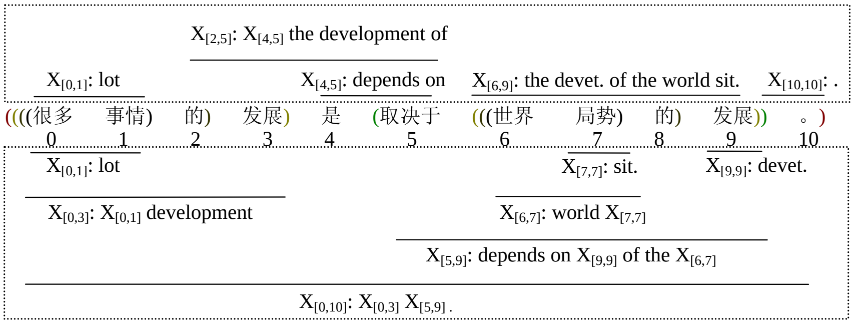

X[2,5]: X[4,5] the development of
X[0,1]: lot
X[4,5]: depends on
X[6,9]: the devet. of the world sit.
X[10,10]: .
((((很多 事情) 的) 发展) 是 (取决于 (((世界 局势) 的) 发展)) 。)
0 1 2 3 4 5 6 7 8 9 10
X[0,1]: lot
X[7,7]: sit.
X[9,9]: devet.
X[0,3]: X[0,1] development
X[6,7]: world X[7,7]
X[5,9]: depends on X[9,9] of the X[6,7]
X[0,10]: X[0,3] X[5,9] .

Figure 3: A translation example of the base HPB system (above) and the system with constraints (below).

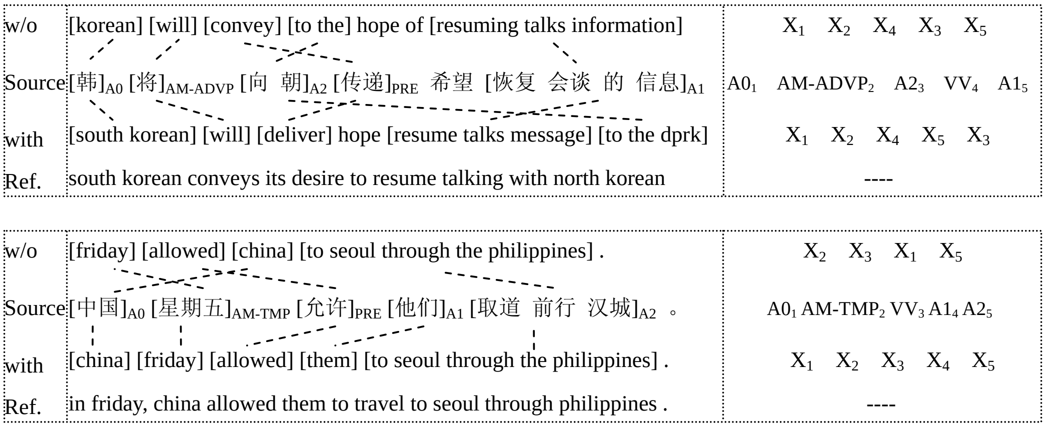

w/o
[korean] [will] [convey] [to the] hope of [resuming talks information]
X1 X2 X4 X3 X5
Source
[韩]A0 [将]AM-ADVP [向 朝]A2 [传递]PRE 希望 [恢复 会谈 的 信息]A1
A0₁ AM-ADVP₂ A2₃ VV₄ A1₅
with
[south korean] [will] [deliver] hope [resume talks message] [to the dprk]
X1 X2 X4 X5 X3
Ref.
south korean conveys its desire to resume talking with north korean
----
w/o
[friday] [allowed] [china] [to seoul through the philippines] .
X2 X3 X1 X5
Source
[中国]A0 [星期五]AM-TMP [允许]PRE [他们]A1 [取道 前行 汉城]A2 。
A0₁ AM-TMP₂ VV₃ A1₄ A2₅
with
[china] [friday] [allowed] [them] [to seoul through the philippines] .
X1 X2 X3 X4 X5
Ref.
in friday, china allowed them to travel to seoul through philippines .

Figure 4: Two translation examples of the system with/without PAS reordering model

short/simple phrase (e.g., *friday*) or a long/complex one (e.g., *when I was 20 years old*), which has impact on its reordering in translation.

6 Related Work

While there has been substantial work on linguistically motivated SMT, we limit ourselves here to several approaches that leverage syntactic constraints yet still allow cross-constituent translations. In terms of tree-based SMT with cross-constituent translations, Cowan et al. (2006) allowed non-constituent sub phrases on the source side and adopted phrase-based translation model for modifiers in clauses. Marcu (2006) and Galley et al. (2006) inserted artificial constituent nodes in parsing tree as to capture useful but non-constituent phrases. The parse tree binarization approach (Wang et al., 2007; Marcu, 2007) and the forest-based approach (Mi et al., 2008) would also cover non-constituent phrases to some extent. Shen et al. (2010) defined well-formed dependency structure to cover uncompleted dependency structure in translation rules. In addition to the fact that the constraints of Shen et al. (2010) and this paper are based on different syntactic perspectives (i.e., dependency structure vs. constituency structure), the major difference is that in this work we don't limit the length of phrases to a fixed maximum size (e.g., 10 in Hiero). Consequently, we obtain some translation rules that are not found in Hiero systems constrained by the length. In terms of (hierarchical) phrase-based SMT with syntactic constraints, particular related to constituent boundaries, Koehn et al. (2003) tested constraints allowing constituent matched phrases only. Chiang (2005) and Cherry (2008) used a soft constraint to award or penalize hypotheses which respect or violate syntactic boundaries. Marton and Resnik (2008) further explored the idea of soft constraints by distinguishing among constituent types. Xiong et al. (2009; 2010) presented models that learn phrase boundaries from aligned dataset.

On the other hand, semantics motivated SMT has also seen an increase in activity recently. Wu and

Fung (2009) re-ordered arguments on the target side translation output, seeking to maximize the cross-lingual match of the semantic frames of the re-ordered translation to that of the source sentence. Liu and Gildea (2010) added two types of semantic role features into a tree-to-string translation model. Although Xiong et al. (2012) and our work are both focusing on source side PAS reordering, our model differs from theirs in two main aspects: 1) we consider reordering not only between an argument and its predicate, but also between two arguments; and 2) our reordering model can naturally model cases of untranslated arguments or predicates.

7 Conclusion

In this paper, we have presented an approach to incorporating syntactic and semantic structures for the HPB translation model. To accommodate the close tie of semantic structures to syntax, we first revisited the idea of hard syntactic constraints, and we demonstrated that hard constraints can, in fact, lead to significant improvement in translation quality when applied to Chiang's HPB framework. Then our PAS reordering model, thanks to the constraints which guided translation phrases in favor of syntactic boundaries, made further improvements by predicting reordering not only between an argument and its predicate, but also between two arguments. In the future work, we will extend the PAS reordering model to include useful context, e.g., the head words and the syntactic categories of arguments.

Acknowledgments

This research was supported in part by the BOLT program of the Defense Advanced Research Projects Agency, Contract No. HR0011-12-C-0015. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the view of DARPA. The authors would like to thank three anonymous reviewers for providing helpful suggestions, and also acknowledge Ke Wu and other CLIP labmates in MT group for useful discussions. We also thank creators of the valuable off-the-shelf NLP packages, such as GIZA++ and Berkeley Parser.

References

Colin Cherry. 2008. Cohesive phrase-based decoding for statistical machine translation. In *Proceedings of ACL-HLT 2008*, pages 72–80.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL 2005*, pages 263–270.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Brooke Cowan, Ivona Kučerová, and Michael Collins. 2006. A discriminative model for tree-to-tree translation. In *Proceedings of EMNLP 2006*, pages 232–241.

Pascale Fung, Zhaojun Wu, Yongsheng Yang, and Dekai Wu. 2006. Automatic learning of Chinese-English semantic structure mapping. In *Proceedings of SLT 2006*, pages 230–233.

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 *Proceedings of ACL-COLING 2006*, pages 961–968.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL 2003*, pages 48–54.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395.

Junhui Li, Guodong Zhou, and Hwee Tou Ng. 2010. Joint syntactic and semantic parsing of Chinese. In *Proceedings of ACL 2010*, pages 1108–1117.

Ding Liu and Daniel Gildea. 2010. Semantic role features for machine translation. In *Proceedings of COLING 2010*, pages 716–724.

Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. 2006. SPMT: Statistical machine translation with syntactified target language phrases. In *Proceedings of EMNLP 2006*, pages 44–52.

Steve DeNeefe; Kevin Knight; Wei Wang; Daniel Marcu. 2007. What can syntax-based mt learn from phrase-based mt? In *Proceedings of EMNLP-CoNLL 2007*, pages 755–763.

Yuval Marton and Philip Resnik. 2008. Soft syntactic constraints for hierarchical phrased-based translation. In *Proceedings of ACL-HLT 2008*, pages 1003–1011.

Haitao Mi, Liang Huang, and Qun Liu. 2008. Forest-based translation. In *Proceedings of ACL-HLT 2008*, pages 192–199.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of ACL 2000*, pages 440–447.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL 2003*, pages 160–167.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of NAACL-HLT 2007*, pages 404–411.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2010. String-to-dependency statistical machine translation. *Computational Linguistics*, 36(4):649–671.

Wei Wang, Kevin Knight, and Daniel Marcu. 2007. Binarizing syntax trees to improve syntax-based machine translation accuracy. In *Proceedings of EMNLP-CoNLL 2007*, pages 746–754.

Dekai Wu and Pascale Fung. 2009. Semantic roles for smt: A hybrid two-pass model. In *Proceedings of NAACL-HLT 2009*, pages 13–16.

Deyi Xiong, Min Zhang, Aiti Aw, and Haizhou Li. 2009. A syntax-driven bracketing model for phrase-based translation. In *Proceedings of ACL-IJCNLP 2009*, pages 315–323.

Deyi Xiong, Min Zhang, and Haizhou Li. 2010. Learning translation boundaries for phrase-based decoding. In *Proceedings of NAACL-HLT 2010*, pages 136–144.

Deyi Xiong, Min Zhang, and Haizhou Li. 2012. Modeling the translation of predicate-argument structure for smt. In *Proceedings of ACL 2012*, pages 902–911.

Nianwen Xue. 2008. Automatic labeling of semantic roles. *Computational Linguistics*, 34(4):225–255.

Tao Zhuang and Chengqing Zong. 2010. Joint inference for bilingual semantic role labeling. In *Proceedings of EMNLP 2010*, pages 304–314.