A Unified Model for Soft Linguistic Reordering Constraints in Statistical Machine Translation

Junhui Li[†] **Yuval Marton**[‡] **Philip Resnik**[†] **Hal Daumé III**[†]
[†]UMIACS, University of Maryland, College Park, MD
{lijunhui, resnik, hal}@umiacs.umd.edu
[‡]Microsoft Corp., City Center Plaza, Bellevue, WA
yumarton@microsoft.com

Abstract

This paper explores a simple and effective unified framework for incorporating soft linguistic reordering constraints into a hierarchical phrase-based translation system: 1) a syntactic reordering model that explores reorderings for context free grammar rules; and 2) a semantic reordering model that focuses on the reordering of predicate-argument structures. We develop novel features based on both models and use them as soft constraints to guide the translation process. Experiments on Chinese-English translation show that the reordering approach can significantly improve a state-of-the-art hierarchical phrase-based translation system. However, the gain achieved by the semantic reordering model is limited in the presence of the syntactic reordering model, and we therefore provide a detailed analysis of the behavior differences between the two.

1 Introduction

Reordering models in statistical machine translation (SMT) model the word order difference when translating from one language to another. The popular distortion or lexicalized reordering models in phrase-based SMT make good local predictions by focusing on reordering on word level, while the synchronous context free grammars in hierarchical phrase-based (HPB) translation models are capable of handling non-local reordering on the translation phrase level. However, reordering, especially without any help of external knowledge, remains a great challenge because an accurate reordering is usually beyond these word level or translation phrase level reordering models' ability. In addition, often these translation models fail to respect linguistically-motivated syntax and semantics. As a result, they tend to produce translations containing both syntactic and semantic reordering confusions. In this paper our goal is to take advantage of syntactic and semantic parsing to improve translation quality. Rather than introducing reordering models on either the word level or the translation phrase level, we propose a unified approach to modeling reordering on the linguistic unit level, e.g., syntactic constituents and semantic roles. The reordering unit falls into multiple granularities, from single words to more complex constituents and semantic roles, and often crosses translation phrases. To show the effectiveness of our reordering models, we integrate both syntactic constituent reordering models and semantic role reordering models into a state-of-the-art HPB system (Chiang, 2007; Dyer et al., 2010). We further contrast it with a stronger baseline, already including fine-grained soft syntactic constraint features (Marton and Resnik, 2008; Chiang et al., 2008). The general ideas, however, are applicable to other translation models, e.g., phrase-based model, as well.

Our syntactic constituent reordering model considers context free grammar (CFG) rules in the source language and predicts the reordering of their elements on the target side, using word alignment information. Due to the fact that a constituent, especially a long one, usually maps into multiple discontinuous blocks in the target language, there is more than one way to describe the monotonicity or swapping patterns; we therefore design two reordering models: one is based on the leftmost aligned target word and the other based on the rightmost target word.

While recently there has also been some encouraging work on incorporating semantic structure (or, more specifically, predicate-argument structure: PAS) reordering in SMT, it is still an open question whether semantic structure reordering

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1123–1133,
Baltimore, Maryland, USA, June 23-25 2014.

strongly overlaps with syntactic structure reordering, since the semantic structure is closely tied to syntax. To this end, we employ the same reordering framework as syntactic constituent reordering and focus on semantic roles in a PAS. We then analyze the differences between the syntactic and semantic features.

The contributions of this paper include the following:

- We introduce novel soft reordering constraints, using syntactic constituents or semantic roles, composed over word alignment information in translation rules used during decoding time;
- We introduce a unified framework to incorporate syntactic and semantic reordering constraints;
- We provide a detailed analysis providing insight into why the semantic reordering model is significantly less effective when syntactic reordering features are also present.

The rest of the paper is organized as follows. Section 2 provides an overview of HPB translation model. Section 3 describes the details of our unified reordering models. Section 4 gives our experimental results and Section 5 discusses the behavior difference between syntactic constituent reordering and semantic role reordering. Section 6 reviews related work and, finally Section 7 concludes the paper.

2 HPB Translation Model: an Overview

In HPB models (Chiang, 2007), 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 $\{\phi_i\}$, such as phrase translation probability $p(\alpha \mid \gamma)$ and its inverse $p(\gamma \mid \alpha)$, the lexical translation probability $p_{lex}(\alpha \mid \gamma)$ and its inverse $p_{lex}(\gamma \mid \alpha)$, 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 log-probability is estimated as:

$$\log P(d) \propto \sum_i \lambda_i \phi_i(d) \qquad (1)$$

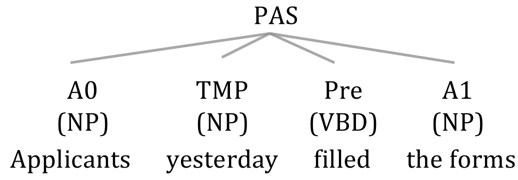

Figure 1: Example of predicate-argument structure.

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

3 Unified Linguistic Reordering Models

As mentioned earlier, the linguistic reordering unit is the syntactic constituent for syntactic reordering, and the semantic role for semantic reordering. The syntactic reordering model takes a CFG rule (e.g., VP $\rightarrow$ VP PP PP) and models the reordering of the constituents on the left hand side by examining their translation or visit order according to the target language. For the semantic reordering model, it takes a PAS and models its reordering on the target side. Figure 1 shows an example of a PAS where the predicate (Pre) has two core arguments (A0 and A1) and one adjunct (TMP). Note that we refer all core arguments, adjuncts, and predicates as semantic roles; thus we say the PAS in Figure 1 has 4 roles. According to the annotation principles in (Chinese) PropBank (Palmer et al., 2005; Xue and Palmer, 2009), all the roles in a PAS map to a corresponding constituent in the parse tree, and these constituents (e.g., NPs and VBD in Figure 1) do not overlap with each other.

Next, we use a CFG rule to describe our syntactic reordering model. Treating the two forms of reorderings in a unified way, the semantic reordering model is obtainable by regarding a PAS as a CFG rule and considering a semantic role as a constituent.

Because the translation of a source constituent might result in multiple discontinuous blocks, there can be several ways to describe or group the reordering patterns. Therefore, we design two general constituent reordering sub-models. One is based on the leftmost aligned word (leftmost reordering model) and the other is based on the rightmost aligned word (rightmost reordering model), as follows. Figure 2 shows the modeling steps for the leftmost reordering model. Figure 2(a) is an example of a CFG rule in the source

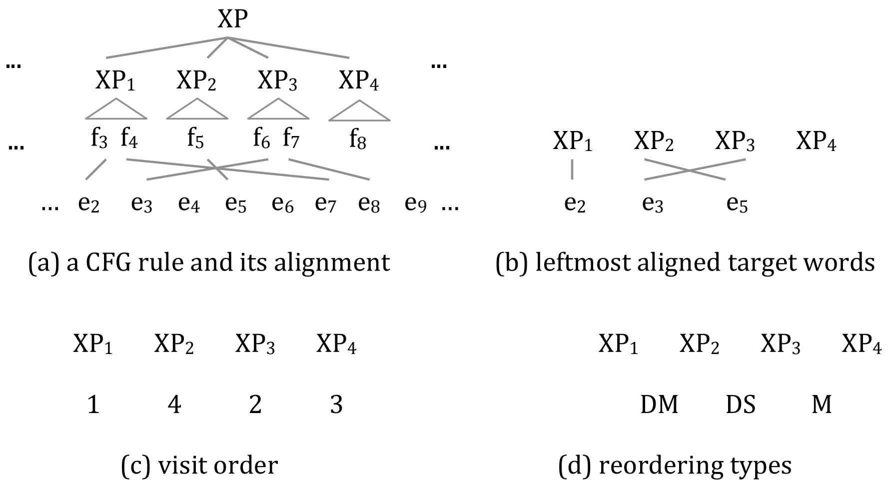

Figure 2: Modeling process illustration for leftmost reordering model.

parse tree and its word alignment links to the target language. Note that constituent XP_4, which covers word f_8, has no alignment. Then for each XP_i, we find the leftmost target word which is aligned to a source word covered by XP_i. Figure 2(b) shows that the leftmost target words for XP_1, XP_2, and XP_3 are e_2, e_5, and e_3, respectively, while XP_4 has no aligned target word. Then we get visit order $V = \{v_i\}$ for $\{XP_i\}$ in the transformation from Figure 2(b) to Figure 2(c), with the following strategies for special cases:

- if the first constituent XP_1 is unaligned, we add a NULL word at the beginning of the target side and link XP_1 to the NULL word;
- if a constituent XP_i ($i > 1$) is unaligned, we add a link to the target word which is aligned to XP_{i-1}, e.g., XP_4 will be linked to e_3; and
- if k constituents $XP_{m_1} \dots XP_{m_k}$ ($m_1 < \dots < m_k$) are linked to the same target word, then $v_{m_i} = v_{m_{i+1}} - 1$, e.g., since XP_3 and XP_4 are both linked to e_3, then $v_3 = v_4 - 1$.

Finally Figure 2(d) converts the visit order $V = \{v_1, \dots v_n\}$ into a sequence of leftmost reordering types $LRT = \{lrt_1, \dots, lrt_{n-1}\}$. For every two adjacent constituents XP_i and XP_{i+1} with corresponding visit order v_i and v_{i+1}, their reordering could be one of the following:

- **Monotone (M)** if $v_{i+1} = v_i + 1$;
- **Discontinuous Monotone (DM)** if $v_{i+1} > v_i + 1$;
- **Swap (S)** if $v_{i+1} = v_i - 1$;
- **Discontinuous Swap (DS)** if $v_{i+1} < v_i - 1$.

Up to this point, we have generated a sequence of leftmost reordering types $LRT = \{lrt_1, \dots, lrt_{n-1}\}$ for a given CFG rule cfg: $XP \rightarrow XP_1 \dots XP_n$. The leftmost reordering model takes the following form:

$$score_{lrt}(cfg) = P_l(lrt_1, \dots, lrt_{n-1} \mid \psi(cfg)) \tag{2}$$

where $\psi(cfg)$ indicates the surrounding context of the CFG. By assuming that any two reordering types in $LRT = \{lrt_1, \dots, lrt_{n-1}\}$ are independent of each other, we reformulate Eq. 2 into:

$$score_{lrt}(cfg) = \prod_{i=1}^{n-1} P_l(lrt_i \mid \psi(cfg)) \tag{3}$$

Similarly, the sequence of rightmost reordering types RRT can be decided for a CFG rule $XP \rightarrow XP_1 \dots XP_n$.

Accordingly, for a PAS pas: $PAS \rightarrow R_1 \dots R_n$, we can obtain its sequences of leftmost and rightmost reordering types by using the same way described above.

3.1 Probability Estimation

In order to predict either the leftmost or rightmost reordering type for two adjacent constituents, we use a maximum entropy classifier to estimate the probability of the reordering type $rt \in \{M, DM, S, DS\}$ as follows:

$$P(rt \mid \psi(cfg)) = \frac{exp\left(\sum_k \theta_k f_k(rt, \psi(cfg))\right)}{\sum_{rt'} exp\left(\sum_k \theta_k f_i(rt', \psi(cfg))\right)} \tag{4}$$

where f_k are binary features, θ_k are the weights of these features. Most of our features f_k are syntax-based. For XP_i and XP_{i+1} in cfg, the features

#Index	Feature
cf1	$\mathcal{L}(XP_i)$ & $\mathcal{L}(XP_{i+1})$ & $\mathcal{L}(XP)$
cf2	for each XP_j $(j < i)$ $\mathcal{L}(XP_i)$ & $\mathcal{L}(XP_{i+1})$ & $\mathcal{L}(XP)$ & $\mathcal{L}(XP_j)$
cf3	for each XP_j $(j > i+1)$ $\mathcal{L}(XP_i)$ & $\mathcal{L}(XP_{i+1})$ & $\mathcal{L}(XP)$ & $\mathcal{L}(XP_j)$
cf4	$\mathcal{L}(XP_i)$ & $\mathcal{L}(XP_{i+1})$ & $\mathcal{P}(XP_i)$
cf5	$\mathcal{L}(XP_i)$ & $\mathcal{L}(XP_{i+1})$ & $\mathcal{H}(XP_i)$
cf6	$\mathcal{L}(XP_i)$ & $\mathcal{L}(XP_{i+1})$ & $\mathcal{P}(XP_{i+1})$
cf7	$\mathcal{L}(XP_i)$ & $\mathcal{L}(XP_{i+1})$ & $\mathcal{H}(XP_{i+1})$
cf8	$\mathcal{L}(XP_i)$ & $\mathcal{L}(XP_{i+1})$ & $\mathcal{S}(XP_i)$
cf9	$\mathcal{L}(XP_i)$ & $\mathcal{L}(XP_{i+1})$ & $\mathcal{S}(XP_{i+1})$
cf10	$\mathcal{L}(XP_i)$ & $\mathcal{L}(XP)$
cf11	$\mathcal{L}(XP_{i+1})$ & $\mathcal{L}(XP)$

Table 1: Features adopted in the syntactic leftmost and rightmost reordering models. $\mathcal{L}(XP)$ returns the syntactic category of *XP*, e.g., NP, VP, PP etc.; $\mathcal{H}(XP)$ returns the head word of XP; $\mathcal{P}(XP)$ returns the POS tagger of the head word; $\mathcal{S}(XP)$ returns the translation status of *XP* on the target language: *un.* if it is untranslated; *cont.* if it is a continuous block; and *discont.* if it maps into multiple discontinuous blocks.

#Index	Feature
rf1	$\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$ & $\mathcal{P}(pas)$ $\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$
rf2	for each R_j $(j < i)$ $\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$ & $\mathcal{R}(R_j)$ & $\mathcal{P}(pas)$ $\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$ & $\mathcal{R}(R_j)$
rf3	for each R_j $(j > i+1)$ $\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$ & $\mathcal{R}(R_j)$ & $\mathcal{P}(pas)$ $\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$ & $\mathcal{R}(R_j)$
rf4	$\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$ & $\mathcal{P}(R_i)$
rf5	$\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$ & $\mathcal{H}(R_i)$
rf6	$\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$ & $\mathcal{L}(R_i)$
rf7	$\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$ & $\mathcal{P}(R_{i+1})$
rf8	$\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$ & $\mathcal{H}(R_{i+1})$
rf9	$\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$ & $\mathcal{L}(R_{i+1})$
rf10	$\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$ & $\mathcal{S}(R_i)$
rf11	$\mathcal{R}(R_i)$ & $\mathcal{R}(R_{i+1})$ & $\mathcal{S}(R_{i+1})$
rf12	$\mathcal{R}(R_i)$ & $\mathcal{P}(pas)$ $\mathcal{R}(R_i)$
rf13	$\mathcal{R}(R_{i+1})$ & $\mathcal{P}(pas)$ $\mathcal{R}(R_{i+1})$

Table 2: Features adopted in the semantic leftmost and rightmost reordering models. $\mathcal{P}(pas)$ returns the predicate content of *pas*; $\mathcal{R}(R)$ returns the role type of *R*, e.g., *Pred, A0, TMP,* etc. For features rf1, rf2, rf3, rf12 and rf13, we include another version which excludes the predicate content $\mathcal{P}(pas)$ for reasons of sparsity.

are aimed to examine which of them should be translated first. Therefore, most features share two common components: the syntactic categories of XP_i and XP_{i+1}. Table 1 shows the features used in syntactic leftmost and rightmost reordering models. Note that we use the same features for both.

Although the semantic reordering model is structured in precisely the same way, we use different feature sets to predict the reordering between two semantic roles. Given the two adjacent roles R_i and R_{i+1} in a PAS *pas*, Table 2 shows the features that are used in the semantic leftmost and rightmost reordering models.

3.2 Integrating into the HPB Model

For models with syntactic reordering, we add two new features (i.e., one for the leftmost reordering model and the other for the rightmost reordering model) into the log-linear translation model in Eq. 1. Unlike the conventional phrase and lexical translation features, whose values are phrase pair-determined and thus can be calculated offline, the value of the reordering features can only be obtained during decoding time, and requires word alignment information as well. Before we present the algorithm integrating the reordering models, we define the following functions by assuming XP_i and XP_{i+1} are the constituent pair of interest in CFG rule *cfg*, H is the translation hypothesis and a is its word alignment:

- $\mathcal{F}_1(w_1, w_2, XP)$: returns *true* if constituent *XP* is within the span from word w_1 to w_2; otherwise returns *false*.
- $\mathcal{F}_2(H, cfg, XP_i, XP_{i+1})$ returns *true* if the reordering of the pair $\langle XP_i, XP_{i+1}\rangle$ in rule *cfg* has not been calculated yet; otherwise returns *false*.
- $\mathcal{F}_3(H, a, XP_i, XP_{i+1})$ returns the leftmost and rightmost reordering types for the constituent pair $\langle XP_i, XP_{i+1}\rangle$, given alignment a, according to Section 3.
- $\mathcal{F}_4(rt, cfg, XP_i, XP_{i+1})$ returns the probability of leftmost reordering type rt for the constituent pair $\langle XP_i, XP_{i+1}\rangle$ in rule *cfg*.
- $\mathcal{F}_5(rt, cfg, XP_i, XP_{i+1})$ returns the probability of rightmost reordering type rt for the constituent pair $\langle XP_i, XP_{i+1}\rangle$ in rule *cfg*.

Algorithm 1 integrates the syntactic leftmost and rightmost reordering models into a CKY-style decoder whenever a new hypothesis is generated. Given a hypothesis H with its alignment a, it traverses all CFG rules in the parse tree and sees if two adjacent constituents are conditioned to trigger the reordering models (lines 2-4). For each pair of constituents, it first extracts its leftmost and rightmost reordering types (line 6) and then gets their respective probabilities returned by the maximum entropy classifiers defined in Section 3.1

Algorithm 1: Integrating the syntactic reordering models into a CKY-style decoder

Input: Sentence f in the source language
Parse tree t of f
All CFG rules $\{cfg\}$ in t
Hypothesis H spanning from word w_1 to w_2
Alignment a of H
Output: Log-Probabilities of the syntactic leftmost and rightmost reordering models
1. set $l\_prob = r_prob = 0.0$
2. **foreach** *cfg* in $\{cfg\}$
3. **foreach** pair XP_i and XP_{i+1} in *cfg*
4. **if** $\mathcal{F}_1$ $(w_1,\, w_2,\, XP_i) = false$ or $\mathcal{F}_1$ $(w_1,\, w_2,\, XP_{i+1}) = false$ or $\mathcal{F}_2$ $(H,\, cfg,\, XP_i,\, XP_{i+1}) = false$
5. **continue**
6. $(l\_type,\, r\_type) = \mathcal{F}_3$ $(H,\, a,\, XP_i,\, XP_{i+1})$
7. l_prob += $\log \mathcal{F}_4$ $(l\_type, cfg, XP_i, XP_{i+1})$
8. r_prob += $\log \mathcal{F}_5$ $(r\_type, cfg, XP_i, XP_{i+1})$
9. **return** $(l\_prob,\, r\_prob)$

(lines 7-8). Then the algorithm returns two log-probabilities of the syntactic reordering models. Note that Function $\mathcal{F}_1$ returns true if hypothesis H fully covers, or fully contains, constituent XP_i, regardless of the reordering type of XP_i. Do not confuse any parsing tag XP_i with the nameless variables X_i in Hiero or cdec rules.

For the semantic reordering models, we also add two new features into the log-linear translation model. To get the two semantic reordering model feature values, we simply use Algorithm 1 and its associated functions from $\mathcal{F}_1$ to $\mathcal{F}_5$ replacing a CFG rule *cfg* with a PAS *pas*, and a constituent XP_i with a semantic role R_i. Algorithm 1 therefore permits a unified treatment of syntactic and PAS-based reordering, even though it is expressed in terms of syntactic reordering here for ease of presentation.

4 Experiments

We have presented our unified approach to incorporating syntactic and semantic soft reordering constraints in an HPB system. In this section, we test its effectiveness in Chinese-English translation.

4.1 Experimental Settings

For training we use 1.6M sentence pairs of the non-UN and non-HK Hansards portions of NIST MT training corpora, segmented with the Stanford segmenter (Tseng et al., 2005). The English data is lowercased, tokenized and aligned with GIZA++ (Och and Ney, 2000) to obtain bidirectional alignments, which are symmetrized using the *grow-diag-final-and* method (Koehn et al., 2003). We train a 4-gram LM on the English side of the corpus with 600M additional words from non-NYT and non-LAT, randomly selected portions of the Gigaword v4 corpus, using modified Kneser-Ney smoothing (Chen and Goodman, 1996). We use the HPB decoder cdec (Dyer et al., 2010), with Mr. Mira (Eidelman et al., 2013), which is a k-best variant of MIRA (Chiang et al., 2008), to tune the parameters of the system.

We use NIST MT 06 dataset (1664 sentence pairs) for tuning, and NIST MT 03, 05, and 08 datasets (919, 1082, and 1357 sentence pairs, respectively) for evaluation.[1] We use BLEU (Papineni et al., 2002) for both tuning and evaluation.

To obtain syntactic parse trees and semantic roles on the tuning and test datasets, we first parse the source sentences with the Berkeley Parser (Petrov and Klein, 2007), trained on the Chinese Treebank 7.0 (Xue et al., 2005). We then pass the parses to a Chinese semantic role labeler (Li et al., 2010), trained on the Chinese PropBank 3.0 (Xue and Palmer, 2009), to annotate semantic roles for all verbal predicates (part-of-speech tag *VV*, *VE*, or *VC*).

Our basic baseline system employs 19 basic features: a language model feature, 7 translation model features, word penalty, unknown word penalty, the glue rule, date, number and 6 pass-through features. Our stronger baseline employs, in addition, the fine-grained syntactic soft constraint features of Marton and Resnik (2008), hereafter MR08. The syntactic soft constraint features include both MR08 exact-matching and cross-boundary constraints (denoted *XP*= and *XP*+). Since the syntactic parses of the tuning and test data contain 29 types of constituent labels and 35 types of POS tags, we have 29 types of *XP*+ features and 64 types of *XP*= features.

4.2 Model Training

To train the syntactic and semantic reordering models, we use a gold alignment dataset.[2] It contains 7,870 sentences with 191,364 Chinese words and 261,399 English words. We first run syn-

[1] http://www.itl.nist.gov/iad/mig//tests/mt

[2] This dataset includes LDC2006E86, and newswire parts of LDC2012T16, LDC2012T20, LDC2012T24, and LDC2013T05. Indeed, the reordering models can also be trained on the MT training data with its automatic alignment. However, our preliminary experiments showed that the reordering models trained on gold alignment yielded higher improvement.

Reordering Type	Syntactic		Semantic	
	l-m	r-m	l-m	r-m
M	73.5	80.6	63.8	67.9
DM	3.9	3.3	14.0	12.0
S	19.5	13.2	13.1	10.7
DS	3.2	3.0	9.1	9.5
#instance	199,234		66,757	

Table 3: Reordering type distribution over the reordering model's training data. Hereafter, l-m and r-m are for leftmost and rightmost, respectively.

tactic parsing and semantic role labeling on the Chinese sentences, then train the models by using MaxEnt toolkit with L1 regularizer (Tsuruoka et al., 2009).[3] Table 3 shows the reordering type distribution over the training data. Interestingly, about 17% of the syntactic instances and 16% of the semantic instances differ in their leftmost and rightmost reordering types, indicating that the leftmost/rightmost distinction is informative. We also see that the number of semantic instances is about 1/3 of that of syntactic instances, but the entropy of the semantic reordering classes is higher, indicating the reordering of semantic roles is harder than that of syntactic constituents.

A deeper examination of the reordering model's training data reveals that some constituent pairs and semantic role pairs have a preference for a specific reordering type (monotone or swap). In order to understand how well the MR08 system respects their reordering preference, we use the gold alignment dataset LDC2006E86, in which the source sentences are from the Chinese Treebank, and thus both the gold parse trees and gold predicate-argument structures are available. Table 4 presents examples comparing the reordering distribution between gold alignment and the output of the MR08 system. For example, the first row shows that based on the gold alignment, for $\langle$*PP,VP*$\rangle$, 16% are in monotone and 76% are in swap reordering. However, our MR08 system outputs 46% of them in monotone and and 50% in swap reordering. Hence, the reordering accuracy for $\langle$*PP,VP*$\rangle$ is 54%. Table 4 also shows that the semantic reordering between core arguments and predicates (e.g., $\langle$*Pred,A1*$\rangle$, $\langle$*A0, Pred*$\rangle$) has a less ambiguous pattern than that between adjuncts and other roles (e.g., $\langle$*LOC,Pred*$\rangle$, $\langle$*A0,TMP*$\rangle$), indicating the higher reordering flexibility of adjuncts.

[3]http://www.logos.ic.i.u-tokyo.ac.jp/∼tsuruoka/maxent/

Const. Pair		Gold		MR08 output		
		M	S	M	S	acc.
PP	VP	16	76	46	50	54
NP	LC	26	74	58	42	50
DNP	NP	24	72	78	19	39
CP	NP	26	67	84	10	33
NP	DEG	39	61	31	69	66
...		...		...		
all		81	13	79	14	80
Role Pair		**Gold**		**MR08 output**		
		M	S	M	S	acc.
Pred	A1	84	6	82	9	72
A0	Pred	82	11	79	8	75
LOC	Pred	17	30	36	25	49
A0	TMP	35	25	61	6	45
TMP	Pred	30	22	49	19	43
...		...		...		
all		63	13	73	9	64

Table 4: Examples of the reordering distribution (%) of gold alignment and the MR08 system output. For simplicity, we only focus on *(M)onotone* and *(S)wap* based on leftmost reordering.

4.3 Translation Experiment Results

Our first group of experiments investigates whether the syntactic reordering models are able to improve translation quality in terms of BLEU. To this end, we respectively add our syntactic reordering models into both the baseline and MR08 systems. The effect is shown in the rows of "+ synreorder" in Table 5. From the table, we have the following two observations.

- Although the HPB model is capable of handling non-local phrase reordering using synchronous context free grammars, both our syntactic leftmost reordering model and rightmost model are still able to achieve improvement over both the baseline and MR08. This suggests that our syntactic reordering features interact well with the MR08 syntactic soft constraints: the XP+ and XP= features focus on a single constituent each, while our reordering features focus on a pair of constituents each.

- There is no clear indication of whether the leftmost reordering model works better than the other. In addition, integrating both the leftmost and rightmost reordering models has limited improvement over a single reordering model.

Our second group of experiments is to validate the semantic reordering models. Results are

System		Tuning	Test			
		MT06	MT03	MT05	MT08	Avg.
Baseline		34.1	36.1	32.3	27.4	31.9
+ syn-reorder	l-m	35.2	36.9‡	33.6‡	28.4‡	33.0
	r-m	35.2	37.2‡	33.7‡	28.6‡	33.2
	both	35.6	37.1‡	33.6‡	28.8‡	33.1
+ sem-reorder	l-m	34.4	36.7‡	33.0‡	27.8†	32.5
	r-m	34.5	36.7‡	33.1‡	27.8‡	32.5
	both	34.5	37.0‡	33.6‡	27.7†	32.8
+syn+sem		35.5	37.3‡	33.7‡	29.0‡	33.3
MR08		35.6	37.4	34.2	28.7	33.4
+ syn-reorder	l-m	36.0	38.2‡	35.0‡	29.2‡	34.1
	r-m	36.0	38.1‡	34.8‡	29.2‡	34.0
	both	35.9	38.2‡	35.3‡	29.5‡	34.3
+ sem-reorder	l-m	35.8	37.6†	34.7‡	28.7	33.7
	r-m	35.8	37.4	34.5†	28.8	33.6
	both	35.8	37.6†	34.7‡	28.8	33.7
+syn+sem		36.1	38.4‡	35.2‡	29.5‡	34.4

Table 5: System performance in BLEU scores. ‡/†: significant over baseline or MR08 at 0.01 / 0.05, respectively, as tested by bootstrap resampling (Koehn, 2004)

shown in the rows of "+ sem-reorder" in Table 5. Here we observe:

- The semantic reordering models also achieve significant gain of 0.8 BLEU on average over the baseline system, demonstrating the effectiveness of PAS-based reordering. However, the gain diminishes to 0.3 BLEU on the MR08 system.
- The syntactic reordering models outperform the semantic reordering models on both the baseline and MR08 systems.

Finally, we integrate both the syntactic and semantic reordering models into the final system. The two models collectively achieve a gain of up to 1.4 BLEU over the baseline and 1.0 BLEU over MR08 on average, which is shown in the rows of "+syn+sem" in Table 5.

5 Discussion

The trend of the results, summarized as performance gain over the baseline and MR08 systems averaged over all test sets, is presented in Table 6. The syntactic reordering models outperform the semantic reordering models, and the gain achieved by the semantic reordering models is limited in the presence of the MR08 syntactic features. In this section, we look at MR08 system and the systems improving it to explore the behavior differences between the two reordering models.

Coverage analysis: Our statistics show that syntactic reordering features (either leftmost or rightmost) are called 24 times per sentence on average. This is compared to only 9 times per sentence for semantic reordering features. This is not surprising since the semantic reordering features are exclusively attached to predicates, and the span set of the semantic roles is a strict subset of the span set of the syntactic constituents; only 22% of syntactic constituents are semantic roles. On average, a sentences has 4 PASs and each PAS contains 3 semantic roles. Of all the semantic role pairs, 44% are in the same CFG rules, indicating that this part of semantic reordering has overlap with syntactic reordering. Therefore, the PAS model has fewer opportunities to influence reordering.

System	Baseline	MR08
+syn-reorder	1.2	0.9
+sem-reorder	0.8	0.3
+ both	1.4	1.0

Table 6: Performance gain in BLEU over baseline and MR08 systems averaged over all test sets.

Reordering accuracy analysis: The reordering type distribution on the reordering model training data in Table 3 suggests that semantic reordering is more difficult than syntactic reordering. To validate this conjecture on our translation test data, we compare the reordering performance among the MR08 system, the improved systems and the maximum entropy classifiers. For the test set, we have four reference translations. We run GIZA++ on the data combination of our translation training data and test data to get the alignment for the test data and each reference translation. Once we have the (semi-)gold alignment, we compute the gold reordering types between two adjacent syntactic constituents or semantic roles. Then we evaluate the automatic reordering outputs generated from both our translation systems and maximum entropy classifiers. Table 7 shows the accuracy averaged over the four gold reordering sets (the four reference translations). It shows that 1) as expected, our classifiers do worse on the harder semantic reordering prediction than syntactic reordering prediction; 2) thanks to the high accuracy obtained by the maxent classifiers, integrating either the syntactic or the semantic reordering constraints results in better reordering performance from both syntactic and semantic perspectives; 3) in terms of the mutual impact, the syntactic reordering models help improving semantic reordering more than the semantic reordering

System	Syntactic		Semantic	
	l-m	r-m	l-m	r-m
MR08	75.0	78.0	66.3	68.5
+syn-reorder	78.4	80.9	69.0	70.2
+sem-reorder	76.0	78.8	70.7	72.7
+both	78.6	81.7	70.6	72.1
Maxent Classifier	80.7	85.6	70.9	73.5

Table 7: Reordering accuracy on four gold sets.

System	Syntactic		Semantic	
	l-m	r-m	l-m	r-m
+syn-reorder	1.2	1.2	-	-
+sem-reorder	-	-	0.7	0.9
+both	1.2	1.0	0.5	0.4

Table 8: Reordering feature weights.

models help improving syntactic reordering; and 4) the rightmost models have a learnability advantage over the leftmost models, achieving higher accuracy across the board.

Feature weight analysis: Table 8 shows the syntactic and semantic reordering feature weights. It shows that the semantic feature weights decrease in the presence of the syntactic features, indicating that the decoder learns to trust semantic features less in the presence of the more accurate syntactic features. This is consistent with our observation that semantic reordering is harder than syntactic reordering, as seen in Tables 3 and 7.

Potential improvement analysis: Table 7 also shows that our current maximum entropy classifiers have room for improvement, especially for semantic reordering. In order to explore the error propagation from the classifiers themselves and explore the upper bound for improvement from the reordering models, we perform an "oracle" study, letting the classifiers be aware of the "gold" reordering type between two syntactic constituents or two semantic roles, and returning a higher probability for the gold reordering type and a smaller one for the others (i.e., we set 0.9 for the gold reordering type, and let the other non-gold three types share 0.1). Again, to get the gold reordering type, we run GIZA++ to get the alignment for tuning/test source sentences and each of four reference translations. We report the averaged performance by using the gold reordering type extracted from the four reference translations. Table 9 compares the performance between the non-oracle and oracle settings. We clearly see that using gold syntactic reordering types significantly improves the performance (e.g., 34.9 vs. 33.4 on average) and there is still some room for improvement by building a better maximum entropy classifiers (e.g., 34.9 vs. 34.3). To our surprise, however, the improvement achieved by gold semantic reordering types is still small (e.g., 33.9 vs. 33.4), suggesting that the potential improvement of semantic reordering models is much more limited. And we again see that the improvement achieved by semantic reordering models is limited in the presence of the syntactic reordering models.

	System	MT 03	MT 05	MT 08	Avg.
Non-Oracle	MR08	37.4	34.2	28.7	33.4
	+syn-reorder	38.2	35.3	29.5	34.3
	+sem-reorder	37.6	34.7	28.8	33.7
	+ both	38.4	35.2	29.5	34.4
Oracle	+syn-reorder	39.2	35.9	29.6	34.9
	+sem-reorder	37.9	34.8	28.9	33.9
	+ both	39.1	36.0	29.8	35.0

Table 9: Performance (BLEU score) comparison between non-oracle and oracle experiments.

6 Related Work

Syntax-based reordering: Some previous work pre-ordered words in the source sentences, so that the word order of source and target sentences is similar. The reordering rules were either manually designed (Collins et al., 2005; Wang et al., 2007; Xu et al., 2009; Lee et al., 2010) or automatically learned (Xia and McCord, 2004; Genzel, 2010; Visweswariah et al., 2010; Khalilov and Sima'an, 2011; Lerner and Petrov, 2013), using syntactic parses. Li et al. (2007) focused on finding the n-best pre-ordered source sentences by predicting the reordering of sibling constituents, while Yang et al. (2012) obtained word order by using a reranking approach to reposition nodes in syntactic parse trees. Both are close to our work; however, our model generates reordering features that are integrated into the log-linear translation model during decoding.

Another approach in previous work added soft constraints as weighted features in the SMT decoder to reward good reorderings and penalize bad ones. Marton and Resnik (2008) employed soft syntactic constraints with weighted binary features and no MaxEnt model. They did not explicitly target reordering (beyond applying constraints on HPB rules). Although employing linguistically motivated labels in SCFG is capable of capturing constituent reorderings (Chiang, 2010; Mylonakis and Sima'an, 2011), the rules are sparser than SCFG with nameless non-terminals (i.e., Xs) and soft constraints. Ge (2010) presented a syntax-driven maximum entropy reordering model that predicted the source word translation order. Gao et al. (2011) employed dependency trees to predict the translation order of a word and its head word. Huang et al. (2013) predicted the translation order of two source words.[4] Our work, which shares this approach, differs from their work primarily in that our syntactic reordering models are based on the constituent level, rather than the word level.

Semantics-based reordering: Semantics-based reordering has also seen an increase in activity recently. In the pre-ordering approach, Wu et al. (2011) automatically learned pre-ordering rules from PAS. In the soft constraint or reordering model approach, Liu and Gildea (2010) modeled the reordering/deletion of source-side semantic roles in a tree-to-string translation model. Xiong et al. (2012) and Li et al. (2013) predicted the translation order between either two arguments or an argument and its predicate. Instead of decomposing a PAS into individual units, Zhai et al. (2013) constructed a classifier for each source side PAS. Finally in the post-processing approach category, Wu and Fung (2009) performed semantic role labeling on translation output and reordered arguments to maximize the cross-lingual match of the semantic frames between the source sentence and the target translation. To our knowledge, their semantic reordering models were PAS-specific. In contrast, our model is universal and can be easily adopted to model the reordering of other linguistic units (e.g., syntactic constituents). Moreover, we have studied the effectiveness of the semantic reordering model in different scenarios.

Non-syntax-based reorderings in HPB: Recently we have also seen work on lexicalized reordering models without syntactic information in HPB (Setiawan et al., 2009; Huck et al., 2013; Nguyen and Vogel, 2013). The non-syntax-based reordering approach models the reordering of translation words/phrases while the syntax-based approach models the reordering of syntactic constituents. Although there are overlaps between translation phrases and syntactic constituents, it is reasonable to think that the two reordering approaches can work together well and even complement each other, as the linguistic patterns they capture differ substantially. Setiawan et al. (2013) modeled the orientation decisions between anchors and two neighboring multi-unit chunks which might cross phrase or rule boundaries. Last, we also note that recent work on non-syntax-based reorderings in (flat) phrase-based models (Cherry, 2013; Feng et al., 2013) can also be potentially adopted to hpb models.

[4]Note that they obtained the translation order of source word pairs by predicting the reordering of adjacent constituents, which was quite close to our work.

7 Conclusion and Future Work

In this paper, we have presented a unified reordering framework to incorporate soft linguistic constraints (of syntactic or semantic nature) into the HPB translation model. The syntactic reordering models take CFG rules and model their reordering on the target side, while the semantic reordering models work with PAS. Experiments on Chinese-English translation show that the reordering approach can significantly improve a state-of-the-art hierarchical phrase-based translation system. We have also discussed the differences between the two linguistic reordering models.

There are many directions in which this work can be continued. First, the syntactic reordering model can be extended to model reordering among constituents that cross CFG rules. Second, although we do not see obvious gain from the semantic reordering model when the syntactic model is adopted, it might be beneficial to further jointly consider the two reordering models, focusing on where each one does well. Third, to better examine the overlap or synergy between our approach and the non-syntax-based reordering approach, we will conduct direct comparisons and combinations with the latter.

Acknowledgments

This research was supported in part by the BOLT program of the Defense Advanced Research Projects Agency, Contract No. HR0012-12-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 comments, and also acknowledge Ke Wu, Vladimir Eidelman, Hua He, Doug Oard, Yuening Hu, Jordan Boyd-Graber, and Jyothi Vinjumur for useful discussions.

References

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of ACL 1996*, pages 310–318.

Colin Cherry. 2013. Improved reordering for phrase-based translation using sparse features. In *Proceedings of HLT-NAACL 2013*, pages 22–31.

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proceedings of EMNLP 2008*, pages 224–233.

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

David Chiang. 2010. Learning to translate with source and target syntax. In *Proceedings of ACL 2010*, pages 1443–1452.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL 2005*, pages 531–540.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of ACL 2010 System Demonstrations*, pages 7–12.

Vladimir Eidelman, Ke Wu, Ferhan Ture, Philip Resnik, and Jimmy Lin. 2013. Mr. mira: Open-source large-margin structured learning on mapreduce. In *Proceedings of ACL 2013 System Demonstrations*, pages 199–204.

Minwei Feng, Jan-Thorsten Peter, and Hermann Ney. 2013. Advancements in reordering models for statistical machine translation. In *Proceedings of ACL 2013*, pages 322–332.

Yang Gao, Philipp Koehn, and Alexandra Birch. 2011. Soft dependency constraints for reordering in hierarchical phrase-based translation. In *Proceedings of EMNLP 2011*, pages 857–868.

Niyu Ge. 2010. A direct syntax-driven reordering model for phrase-based machine translation. In *Proceedings of HLT-NAACL 2010*, pages 849–857.

Dmitriy Genzel. 2010. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of COLING 2010*, pages 376–384.

Zhongqiang Huang, Jacob Devlin, and Rabih Zbib. 2013. Factored soft source syntactic constraints for hierarchical machine translation. In *Proceedings of EMNLP 2013*, pages 556–566.

Matthias Huck, Joern Wuebker, Felix Rietig, and Hermann Ney. 2013. A phrase orientation model for hierarchical machine translation. In *Proceedings of WMT 2013*, pages 452–463.

Maxim Khalilov and Khalil Sima'an. 2011. Context-sensitive syntactic source-reordering by statistical transduction. In *Proceedings of IJCNLP 2011*, pages 38–46.

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

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

Young-Suk Lee, Bing Zhao, and Xiaoqian Luo. 2010. Constituent reordering and syntax models for English-to-Japanese statistical machine translation. In *Proceedings of COLING 2010*, pages 626–634.

Uri Lerner and Slav Petrov. 2013. Source-side classifier preordering for machine translation. In *Proceedings of EMNLP 2013*, pages 513–523.

Chi-Ho Li, Minghui Li, Dongdong Zhang, Mu Li, Ming Zhou, and Yi Guan. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of ACL 2007*, pages 720–727.

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

Junhui Li, Philip Resnik, and Hal Daumé III. 2013. Modeling syntactic and semantic structures in hierarchical phrase-based translation. In *Proceedings of HLT-NAACL 2013*, pages 540–549.

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

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

Markos Mylonakis and Khalil Sima'an. 2011. Learning hierarchical translation structure with linguistic annotations. In *Proceedings of ACL 2011*, pages 642–652.

ThuyLinh Nguyen and Stephan Vogel. 2013. Integrating phrase-based reordering features into a chart-based decoder for machine translation. In *Proceedings of ACL 2013*, pages 1587–1596.

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

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318.

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

Hendra Setiawan, Min Yen Kan, Haizhou Li, and Philip Resnik. 2009. Topological ordering of function words in hierarchical phrase-based translation. In *Proceedings of ACL-IJCNLP 2009*, pages 324–332.

Hendra Setiawan, Bowen Zhou, Bing Xiang, and Libin Shen. 2013. Two-neighbor orientation model with cross-boundary global contexts. In *Proceedings of ACL 2013*, pages 1264–1274.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168–171.

Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2009. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of ACL-IJCNLP 2009*, pages 477–485.

Karthik Visweswariah, Jiri Navratil, Jeffrey Sorensen, Vijil Chenthamarakshan, and Nandakishore Kambhatla. 2010. Syntax based reordering with automatically derived rules for improved statistical machine translation. In *Proceedings of COLING 2010*, pages 1119–1127.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of EMNLP 2007*, pages 737–745.

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

Xianchao Wu, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2011. Extracting pre-ordering rules from predicate-argument structures. In *Proceedings of IJCNLP 2011*, pages 29–37.

Fei Xia and Michael McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of COLING 2004*, pages 508–514.

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.

Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. 2009. Using a dependency parser to improve smt for subject-object-verb languages. In *Proceedings of HLT-NAACL 2009*, pages 245–253.

Nianwen Xue and Martha Palmer. 2009. Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, 15(1):143–172.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Nan Yang, Mu Li, Dongdong Zhang, and Nenghai Yu. 2012. A ranking-based approach to word reordering for statistical machine translation. In *Proceedings of ACL 2012*, pages 912–920.

Feifei Zhai, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2013. Handling ambiguities of bilingual predicate-argument structures for statistical machine translation. In *Proceedings of ACL 2013*, pages 1127–1136.