Using Syntactic Head Information in Hierarchical Phrase-Based Translation

Junhui Li Zhaopeng Tu[†] Guodong Zhou[‡] Josef van Genabith
Centre for Next Generation Localisation
School of Computing, Dublin City University
[†] Key Lab. of Intelligent Info. Processing
Institute of Computing Technology, Chinese Academy of Sciences
[‡]School of Computer Science and Technology
Soochow University, China
{jli,josef}@computing.dcu.ie
tuzhaopeng@ict.ac.cn gdzhou@suda.edu.cn

Abstract

Chiang's hierarchical phrase-based (HPB) translation model advances the state-of-the-art in statistical machine translation by expanding conventional phrases to hierarchical phrases – phrases that contain sub-phrases. However, the original HPB model is prone to overgeneration due to lack of linguistic knowledge: the grammar may suggest more derivations than appropriate, many of which may lead to ungrammatical translations. On the other hand, limitations of glue grammar rules in the original HPB model may actually prevent systems from considering some reasonable derivations. This paper presents a simple but effective translation model, called the Head-Driven HPB (HD-HPB) model, which incorporates head information in translation rules to better capture syntax-driven information in a derivation. In addition, unlike the original glue rules, the HD-HPB model allows improved reordering between any two neighboring non-terminals to explore a larger reordering search space. An extensive set of experiments on Chinese-English translation on four NIST MT test sets, using both a small and a large training set, show that our HD-HPB model consistently and statistically significantly outperforms Chiang's model as well as a source side SAMT-style model.

1 Introduction

Chiang's hierarchical phrase-based (HPB) translation model utilizes synchronous context free grammar (SCFG) for translation derivation (Chiang, 2005; Chiang, 2007) and has been widely adopted in statistical machine translation (SMT). Typically, such models define two types of translation rules: hierarchical (translation) rules which consist of both terminals and non-terminals, and glue (grammar) rules which combine translated phrases in a monotone fashion. However, due to lack of linguistic knowledge, Chiang's HPB model contains only one type of non-terminal symbol X, often making it difficult to select the most appropriate translation rules.[1]

One important research question is therefore how to refine the non-terminal category *X* using linguistically motivated information: Zollmann and Venugopal (2006) (SAMT) e.g. use (partial) syntactic categories derived from CFG trees while Zollmann and Vogel (2011) use word tags, generated by either POS analysis or unsupervised word class induction. Almaghout et al. (2011) employ CCG-based supertags. Mylonakis and Sima'an (2011) use linguistic information of various granularities such as *Phrase-Pair*, *Constituent*, *Concatenation of Constituents*, and *Partial Constituents*, where applicable.

By contrast, and inspired by previous work in parsing (Charniak, 2000; Collins, 2003), our Head-Driven HPB (HD-HPB) model is based on the intuition that linguistic heads provide important information about a constituent or distributionally defined fragment, as in HPB. We identify heads using linguistically motivated dependency parsing, and use head information to refine *X*.

Furthermore, Chiang's HPB model suffers from limited phrase reordering by combining translated

[1] Another non-terminal symbol S is used in glue rules.

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 232–242,
Montréal, Canada, June 7-8, 2012.

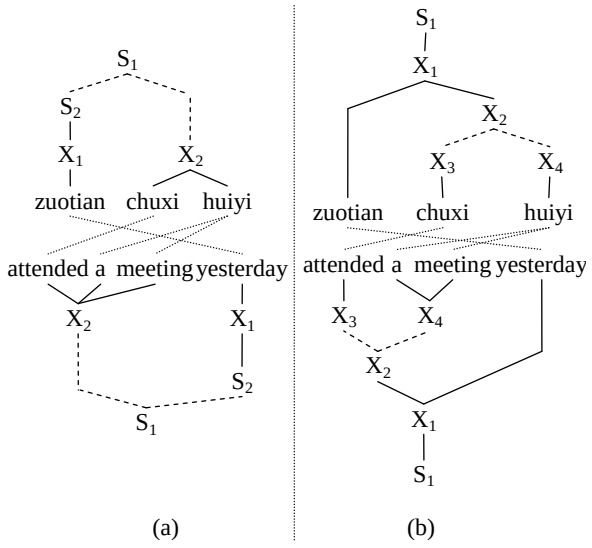

Figure 1: Example of derivations disallowed in Chiang's HPB model. The rules with dotted lines are not covered in Chiang's model.

phrases in a monotonic way with glue rules. In addition, once a glue rule is adopted, it requires all rules above it to be glue rules. For example, given a Chinese-English sentence pair (昨天/*zuotian*$_1$ 出席/*chuxi*$_2$ 会议/*huiyi*$_3$, *Attended*$_2$ a_3 *meeting*$_3$ *yesterday*$_1$), a correct translation is impossible via HPB derivations in Figure 1. For the derivation in Figure 1(a), swap reordering in the glue rule (i.e., $S_1 \rightarrow \langle S_2X_2,\ X_2S_2\rangle$) is disallowed and, even if such a swap reordering is available, it lacks useful information for rule selection. For the derivation in Figure 1(b), the combination of two non-terminals (i.e., $X_2 \rightarrow \langle X_3X_4,\ X_3X_4\rangle$) is disallowed to form a new non-terminal which in turn is a sub-phrase of a hierarchical rule. These limitations prevent traditional HPB systems from even considering some reasonable derivations.

To tackle the problem of glue rules, He (2010) extended the HPB model by using bracketing transduction grammar (Wu, 1996) instead of the monotone glue rules, and trained an extra classifier for glue rules to predict reorderings of neighboring phrases. By contrast, our HD-HPB model refines the non-terminal symbol X with syntactic head information and provides flexible reordering rules, including swap, which can mix freely with hierarchical translation rules for better interleaving of translation and reordering in translation derivations.

Different from the soft constraint modeling adopted in (Chan et al., 2007; Marton and Resnik, 2008; Shen et al., 2009; He et al., 2010; Huang et al., 2010; Gao et al., 2011), our approach encodes syntactic information in translation rules. However, the two approaches are not mutually exclusive, as we could also include a set of syntax-driven features into our translation model. Our approach maintains the advantages of Chiang's HPB model while at the same time incorporating head information and flexible reordering in a derivation in a natural way. Experiments on Chinese-English translation using four NIST MT test sets show that our HD-HPB model significantly outperforms Chiang's HPB as well as a SAMT-style refined version of HPB.

The paper is structured as follows: Section 2 describes the synchronous context-free grammar (SCFG) in our HD-HPB translation model. Section 3 presents our model and features, followed by the decoding algorithm in Section 4. We report experimental results in Section 5. Finally we conclude in Section 6.

2 Head-Driven HPB Translation Model

Like Chiang (2005) and Chiang (2007), our HD-HPB translation model adopts a synchronous context free grammar, a rewriting system which generates source and target side string pairs simultaneously using a context-free grammar. In particular, each synchronous rule rewrites a non-terminal into a pair of strings, s and t, where s (or t) contains terminals and non-terminals from the source (or target) language and there is a one-to-one correspondence between the non-terminal symbols on both sides.

A good and informative inventory of non-terminal symbols is always important, especially for a successful SCFG-based translation model. Instead of collapsing all non-terminals in the source language into a single symbol X as in Chiang (2007), ideally non-terminals should capture important information of the word sequences they cover to be able to properly discriminate between similar and different word sequences during translation. This motivates our approach to provide syntax-enriched non-terminal symbols. Given a word sequence f_j^i from position i to position j, we refine the non-terminal symbol X to reflect some of the internal syntactic structure of

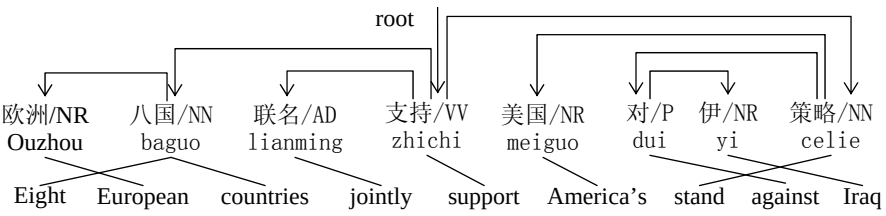

Figure 2: An example word alignment for a Chinese-English sentence pair with the dependency parse tree for the Chinese sentence. Here, each Chinese word is attached with its POS tag and Pinyin.

the word sequence covered by X. A correct translation rule selection therefore not only maps terminals into terminals, but is both constrained and guided by syntactic information in the non-terminals. At the same time, it is not clear whether an "ideal" approach that captures a full syntactic analysis of the string fragment covered by a non-terminal is feasible: the diversity of syntactic structures could make training impossible and lead to serious data sparseness issues. As a compromise, given a word sequence f_j^i, we first find **heads** and then concatenate the POS tags of these heads as f_j^i's non-terminal symbol.[2] Our approach is guided by the intuition that linguistic heads provide important information about a constituent or distributionally defined fragment, as in HPB. Specifically, we adopt dependency structure to derive heads, which are defined as:

Definition 1. *For word sequence f_j^i, word $f_k\,(i \leq k \leq j)$ is regarded as a **head** if it is dominated by a word outside of this sequence.*

Note that this definition (i) allows for a word sequence to have one or more heads (largely due to the fact that a word sequence is not necessarily linguistically constrained) and (ii) ensures that heads are always the highest heads in the sequence from a dependency structure perspective. For example, the word sequence *ouzhou baguo lianming* in Figure 2 has two heads (i.e., *baguo* and *lianming*, *ouzhou* is not a head of this sequence since its headword *baguo* falls within this sequence) and the non-terminal corresponding to the sequence is thus labeled as *NN-AD*. It is worth noting that in this paper we only refine non-terminal X on the source side to head-informed ones, while still using X on the target side.

[2]Note that instead of POS tags, it is also possible to use other types of syntactic information associated with heads to refine non-terminal symbols (Section 5.5.2).

In our HD-HPB model, the SCFG is defined as a tuple $\langle \Sigma, N, \Delta, \Lambda, \Re \rangle$, where Σ is a set of source language terminals, N is a set of non-terminals categorizing terminals in Σ, Δ is a set of target language terminals, Λ is a set of non-terminals categorizing terminals in Δ, and $\Re$ is a set of translation rules. A rule γ in $\Re$ is in the form of $\langle P_s \rightarrow s, P_t \rightarrow t, \phi \rangle$, where:

- $P_s \in N$ and $P_t \in \Lambda$;
- $s \in (\Sigma \cup N)^+$ and $t \in (\Delta \cup \Lambda)^+$
- ϕ is a bijection between non-terminals in s and t.

According to the occurrence of terminals in s and t, we group the rules in the HD-HPB model into two categories: head-driven hierarchical rules (HD-HRs) and non-terminal reordering rules (NRRs), where the former have at least one terminal on both source and target sides and the later have no terminals. For rule extraction, we first identify *initial phrase pairs* on word-aligned sentence pairs by using the same criterion as most phrase-based translation models (Och and Ney, 2004) and Chiang's HPB model (Chiang, 2005; Chiang, 2007). We extract HD-HRs and NRRs based on initial phrase pairs, respectively.

2.1 HD-HRs: Head-Driven Hierarchical Rules

As mentioned, a HD-HR has at least one terminal on both source and target sides. This is the same as the hierarchical rules defined in Chiang's HPB model (Chiang, 2007), except that we use head POS-informed non-terminal symbols in the source language. We look for initial phrase pairs that contain other phrases and then replace sub-phrases with their corresponding non-terminal symbols. Given the word alignment as shown in Figure 2, Table 1 demonstrates the difference between hierarchical rules in Chiang (2007) and HD-HRs defined here.

phrase pairs	hierarchical rule	head-driven hierarchical rule
celie, stand	X→celie, stand	NN→celie, X→stand
dui yi celie$_1$, stand$_1$ against Iraq	X→dui yi X$_1$, X$_1$ against Iraq	NN→dui yi NN$_1$, X→X$_1$ against Iraq
zhichi meiguo, support America's	X→zhichi meiguo, support America's	VV-NR→zhichi meiguo, X→support America's
zhichi meiguo$_1$ dui yi celie$_2$, support America's$_1$ stand$_2$ against Iraq	X→X$_1$ dui yi X$_2$, X$_1$ X$_2$ against Iraq	VV→VV-NR$_1$ dui yi NN$_2$, X→X$_1$ X$_2$ against Iraq

Table 1: Comparison of hierarchical rules in Chiang (2007) and HD-HRs. Indexed underlines indicate sub-phrases and corresponding non-terminal symbols. The non-terminals in HD-HRs (e.g., NN, VV, VV-NR) capture the head(s) POS tags of the corresponding word sequence in the source language.

Similar to Chiang's HPB model, our HD-HPB model will result in a large number of rules causing problems in decoding. To alleviate these problems, we filter our HD-HRs according to the same constraints as described in Chiang (2007). Moreover, we discard rules that have non-terminals with more than four heads.

2.2 NRRs: Non-terminal Reordering Rules

NRRs are translation rules without terminals. Given an initial phrase pair $\left\langle f_j^i, e_{j^*}^{i^*} \right\rangle$, we check all other initial phrase pairs $\left\langle f_l^k, e_{l^*}^{k^*} \right\rangle$ which satisfy $k = j+1$ (i.e., phrase f_l^k is located immediately to the right of f_j^i in the source language). For their target side translations, there are four possible positional relationships: monotone, discontinuous monotone, swap, and discontinuous swap. In order to differentiate non-terminals from those in the target language (i.e., X), we use Y as a variable for non-terminals in the source language, and obtain four types of NRRs:

- Monotone $\langle Y \rightarrow Y_1 Y_2, X \rightarrow X_1 X_2 \rangle$;
- Discontinuous monotone $\langle Y \rightarrow Y_1 Y_2, X \rightarrow X_1 \ldots X_2 \rangle$;
- Swap $\langle Y \rightarrow Y_1 Y_2, X \rightarrow X_2 X_1 \rangle$;
- Discontinuous swap $\langle Y \rightarrow Y_1 Y_2, X \rightarrow X_2 \ldots X_1 \rangle$.

For example in Figure 2, the NRR for initial phrase pairs $\langle$*zhichi meiguo*, *support America's*$\rangle$ and $\langle$*dui yi celie*, *stand against Iraq*$\rangle$ would be $\langle VV \rightarrow VV\text{-}NR_1\ NN_2,\ X \rightarrow X_1\ X_2 \rangle$.

Merging two neighboring non-terminals into a single non-terminal, NRRs enable the translation model to explore a wider search space. During training, we extract four types of NRRs and calculate probabilities for each type. To speed up decoding, we currently (i) only use monotone and swap NRRs and (ii) limit the number of non-terminals in a NRR to 2.

3 Log-linear Model and Features

Following Och and Ney (2002), we depart from the traditional noisy-channel approach and use a general log-linear model. Let d be a derivation from sentence f in the source language to sentence e in the target language. The probability of d is defined as:

$$P(d) \propto \prod_i \emptyset_i(d)^{\lambda_i} \quad (1)$$

where $\emptyset_i$ are features defined on derivations and λ_i are feature weights. In particular, we use a feature set analogous to the default feature set of Chiang (2007), which includes:

- $P_{hd\text{-}hr}(t|s)$ and $P_{hd\text{-}hr}(s|t)$, translation probabilities for HD-HRs;
- $P_{lex}(t|s)$ and $P_{lex}(s|t)$, lexical translation probabilities for HD-HRs;
- $Pty_{hd\text{-}hr} = exp(-1)$, rule penalty for HD-HRs;
- $P_{nrr}(t|s)$, translation probability for NRRs;
- $Pty_{nrr} = exp(-1)$, rule penalty for NRRs;
- $P_{lm}(e)$, language model;
- $Pty_{word}(e) = exp(-|e|)$, word penalty.

Algorithm 1: Decoding Algorithm

Input: Sentence f_n^1 in the source language
Dependency structure of f_n^1
HD-HR rule set *HDHR*
NRR rule set *NRR*
Initial phrase length K
Output: Best derivation d^*
1. set *chart[i, j]=NIL* $(1 \leq i \leq j \leq n)$;
2. **for** l from 1 to n **do**
3. **for** all i, j such that $j - i = l$ **do**
4. **if** $l \leq K$ **do**
5. **for** all derivations d derived from *HDHR* spanning from i to j **do**
6. add d into $chart[i, j]$
7. **for** all derivations d derived from *NRR* spanning from i to j **do**
8. add d into $chart[i, j]$
9. set d^* as the top derivation of $chart[1, n]$
10. **return** d^*

It is worth pointing out that we define translation probabilities for NRRs only for the direction from source language to target language, although translation probabilities for HD-HRs are defined for both directions. This is mostly due to the fact that a NRR excludes terminals and has only two options on the target side (i.e., either $X \rightarrow X_1X_2$ or $X \rightarrow X_2X_1$).

4 Decoding

Our decoder is based on CKY-style chart parsing with beam search. Given an input sentence f, it finds a sentence e in the target language derived from the best derivation d^* among all possible derivations D:

$$d^* = \arg\max_{d \in D} P(D) \quad (2)$$

Algorithm 1 presents the decoding process. Given a source sentence, it searches for the best derivation bottom-up. For a source span $[i, j]$, it applies both types of HD-HRs and NRRs. However, HD-HRs are only applied to generate derivations spanning no more than K words – the initial phrase length limit used in training to extract HD-HRs – while NRRs are applied to derivations spanning any length. Unlike in Chiang (2007), it is possible for a non-terminal generated by a NRR to be included afterwards by a HD-HR or another NRR. Similar to Chiang (2007) in generating k-best derivations from i to j, we make use of cube pruning (Huang and Chiang, 2005) with an integrated language model for each derivation.

5 Experiments

We evaluate the performance of our HD-HPB model and compare it with our implementation of Chiang's HPB model (Chiang, 2007), a source-side SAMT-style refined version of HPB (SAMT-HPB), and the Moses implementation of HPB. For fair comparison, we adopt the same parameter settings for HD-HPB, HPB and SAMT-HPB systems, including initial phrase length (as 10) in training, the maximum number of non-terminals (as 2) in translation rules, maximum number of non-terminals plus terminals (as 5) on the source, prohibition of non-terminals to be adjacent on the source, beam threshold β (as 10^{-5}) (to discard derivations with a score worse than β times the best score in the same chart cell), beam size b (as 200) (i.e. each chart cell contains at most b derivations). For Moses HPB, we use "grow-diag-final-and" to obtain symmetric word alignments, 10 for the maximum phrase length, and the recommended default values for all other parameters.

5.1 Experimental Settings

To examine the efficacy of our approach on training datasets of different scales, we first train translation models on a small-sized corpus, and then scale to a larger one. We use the 2002 NIST MT evaluation test data (878 sentence pairs) as the development data, and the 2003, 2004, 2005, 2006-news NIST MT evaluation test data (919, 1788, 1082, and 616 sentence pairs, respectively) as the test data. To find heads, we parse the source sentences with the Berkeley Parser[3] (Petrov and Klein, 2007) trained on Chinese TreeBank 6.0 and use the Penn2Malt toolkit[4] to obtain dependency structures.

We obtain the word alignments by running GIZA++ (Och and Ney, 2000) on the corpus in both directions, 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

[3] http://code.google.com/p/berkeleyparser/
[4] http://w3.msi.vxu.se/~nivre/research/Penn2Malt.html/

weights on the development data.

For evaluation, the NIST BLEU script (version 12) with the default settings is used to calculate the NIST and the 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 bootstrap approach (Koehn, 2004). In this paper, ' ** ' and ' * ' denote p-values less than 0.01 and in-between [0.01, 0.05), respectively.

5.2 Results on Small Data

To test the HD-HPB models, we firstly carried out experiments using the FBIS corpus as training data, which contains ˜240K sentence pairs. Table 2 lists the rule table sizes. The full rule table size (including HD-HRs and NRRs) of our HD-HPB model is about 1.5 times that of Chiang's, largely due to refining the non-terminal symbol X in Chiang's model into head-informed ones in our model. It is also unsurprising, that the test set-filtered rule table size of our model is only about 0.8 times that of Chiang's: this is due to the fact that some of the refined translation rule patterns required by the test set are unattested in the training data. Furthermore, the rule table size of NRRs is much smaller than that of HD-HRs since a NRR contains only two non-terminals. Table 3 lists the translation performance with NIST and BLEU scores. Note that our re-implementation of Chiang's original HPB model performs on a par with Moses HPB. Table 3 shows that our HD-HPB model significantly outperforms Chiang's HPB model with an average improvement of 1.32 in BLEU and 0.16 in NIST (and similar improvements over Moses HPB).

Although HD-HPB has small size of phrase tables compared to HPB, it still consumes more time in decoding (e.g., 15.1 vs. 11.0), mostly due to the flexible reordering of NRRs.

5.3 Results on Large Data

We also conduct experiments on larger training data with ˜1.5M sentence pairs from the LDC dataset.[5] Table 4 lists the rule table sizes and Table 5 presents translation performance with NIST and BLEU scores. It shows that our HD-HPB model consistently outperforms Chiang's HPB model with an average improvement of 1.91 in BLEU and 0.35 in NIST (similar for Moses HPB). Compared to the improvement achieved on the small data, it is encouraging to see that our HD-HPB model benefits more from larger training data with little adverse effect on decoding time which increases only slightly from 15.1 to 16.6 seconds per sentence.

5.4 Comparison with SAMT-HPB

Comparing the performance of SAMT-HPB with regular HPB in Table 3 and Table 5, it is interesting to see that in general the SAMT-style approach leads to a deterioration of translation performance for the small training set (e.g., 30.09 for SAMT-HPB vs. 30.64 for HPB) while it comes into its own for the large training set (e.g., 33.54 for SAMT-HPB vs. 32.95 for HPB), indicating that the SAMT-style approach is more prone to data sparseness than HPB (or, indeed, HD-HPB).

Comparing the performance of SAMT-HPB with HD-HPB, shows that our head-driven non-terminal refining approach consistently outperforms the SAMT-style approach on an extensive set of experiments (for each test set $p < 0.01$), indicating that head information is more effective than (partial) CFG categories. To make the comparison fair, it is important to note that our implementation of source-side SAMT-HPB includes the same sophisticated non-terminal re-ordering NRR rules as HD-HPB (Section 2.2). Thus the performance differences reported here are not due to different reordering capabilities, but to the discriminative impact of the head information in HD-HPB over SAMT-style annotation. Taking *lianming zhichi* in Figure 2 as an example, HD-HPB labels the span *VV*, as *lianming* is dominated by *zhichi*, effecively ignoring *lianming* in the translation rule, while the SAMT label is *ADVP:AD+VV*[6] which is more susceptible to data sparsity (Table 2 and Table 4). In addition, SAMT resorts to *X* if a text span fails to satisify pre-defined categories. Examining initial phrases extracted from the SAMT training data shows that 28% of them are labeled as *X*. Finally, for Chinese syntactic analy-

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

[6] The constituency structure for *lianming zhichi* is *(VP (ADVP (AD lianming)) (VP (VV zhichi) ...)).*

System	Total Rules	MT 03	MT 04	MT 05	MT 06	Avg.
HPB	39.6M	2.8M	4.7M	3.3M	3.0M	3.4M
HD-HPB	59.5/0.6M	1.9/0.1M	3.4/0.2M	2.3/0.2M	2.0/0.1M	2.4/0.2M
SAMT-HPB	70.1/0.4M	2.2/0.2M	4.0/0.2M	2.7/0.2M	2.3/0.2M	2.8/0.2M

Table 2: Rule table sizes of different models trained on small data. Note: 1) SAMT-HPB indicates our HD-HPB model with the non-terminal scheme of Zollmann and Venugopal (2006); 2) For HD-HPB and SAMT-HPB, the rule sizes separated by / indicate HD-HRs and NRRs, respectively; 2) Except for "Total Rules", the figures correspond to rules filtered on the corresponding test set.

System	MT 03		MT 04		MT 05		MT 06		Avg.		Time
	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	
Moses HPB	7.377	29.67	8.209	33.60	7.571	29.49	6.773	28.90	7.483	30.42	NA
HPB	8.137	29.75	9.050	34.06	8.264	30.09	7.788	28.64	8.310	30.64	11.0
HD-HPB	**8.308**	**31.01****	**9.211**	**35.11****	**8.426**	**31.57****	**7.930**	**30.15****	**8.469**	**31.96**	15.1
SAMT-HPB	7.886	29.14*	8.703	33.32**	7.961	29.49*	7.307	28.41	7.964	30.09	17.3
HD-HR+Glue	7.966	29.51	8.826	33.68	8.116	29.84	7.474	28.51	8.095	30.39	5.4

Table 3: NIST and BLEU (%) scores of different models trained on small data. Note: 1) HD-HR+Glue indicates our HD-HPB model replacing NRRs with glue rules; 2) Significance tests for Moses HPB, HD-HPB, SAMT-HPB and HD-HR+Glue are done against HPB.

System	Total Rules	MT 03	MT 04	MT 05	MT 06	Avg.
HPB	206.8M	11.3M	17.6M	12.9M	10.4M	13.0M
HD-HPB	318.6/2.3M	7.3/0.3M	12.2/0.4M	8.5/0.3M	6.7/0.2M	8.7/0.3M
SAMT-HPB	371.0/1.1M	8.6/0.3M	14.3/0.4M	10.1/0.3M	7.9/0.3M	10.2/0.3M

Table 4: Rule table sizes of different models trained on large data.

System	MT 03		MT 04		MT 05		MT 06		Avg.		Time
	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	
Moses HPB	7.914	32.94*	8.429	35.16	7.962	32.18	6.483	29.88*	7.697	32.54	NA
HPB	8.583	33.59	9.114	35.39	8.465	32.20	7.532	30.60	8.423	32.95	13.7
HD-HPB	**8.885**	**35.50****	**9.494**	**37.61****	**8.871**	**34.56****	**7.839**	**31.78****	**8.772**	**34.86**	16.6
SAMT-HPB	8.644	34.07	9.245	36.52**	8.618	32.90*	7.543	30.66	8.493	33.54	19.1
HD-HR+Glue	8.831	34.58**	9.435	36.55**	8.821	33.84**	7.863	31.06	8.737	34.01	6.7

Table 5: NIST and BLEU (%) scores of different models trained on large data. Note: System labels and significance testing as in Table 3.

sis, dependency structure is more reliable than constituency structure. Moreover, SAMT-HPB takes more time in decoding than HD-HPB due to larger phrase tables.

5.5 Discussion

5.5.1 Individual Contribution of HD-HRs and NRRs

Examining translation output shows that on average each sentence employs 16.6/5.2 HD-HRs/NRRs in our HD-HPB model, compared to 15.9/3.6 hierarchical rules/glue rules in Chiang's model, providing further indication of the importance of NRRs in translation. In order to separate out the individual contributions of the novel HD-HRs and NRRs, we carry out an additional experiment (HD-HR+Glue) using HD-HRs with monotonic glue rules only (adjusted to refined rule labels, but effectively switching off the extra reordering power of full NRRs) both on the small and the large datasets, with interesting results: Table 3 (HD-HR+Glue) shows that for the small training set most of the improvement of our full HD-HPB model comes from the NRRs, as RR+Glue performs on the same level as Chiang's original and Moses HPB (the differences are not statistically significant), perhaps indicating sparseness for the refined HD-HRs given the small training set. Table 5 shows that for the large training set, HD-HRs come into their own: on average more than half of the improvement over HPB (Chiang and Moses) comes from the refined HD-HRs, the rest from NRRs.

It is not surprising that compared to the others HD-HR+Glue takes much less time in decoding. This is due to the fact that 1) compared to HPB, the refined translation rule patterns on the source side have fewer entries in phrase table; 2) compared to HD-HPB, HD-HR+Glue switches off the extra reordering of NRRs. The decoding time for HD-HPB and HD-HR+Glue suggests that NRRs are more than doubling the time required to decode.

5.5.2 Different Head Label Sets

Examining initial phrases extracted from the large size training data shows that there are 63K types of refined non-terminals with respect to 33 types of POS tags. Considering the sparseness in translation rules caused by this comparatively detained POS tag set, we carry out an experiment with a reduced set of non-terminal types by using a less granular POS tag set (C-HPB). Moreover, due to the fact that concatenation of POS tags of heads mostly captures internal structure of a text span, it is interesting to examine the effect of other syntactic labels, in particular dependency labels, to try to better capture the impact of the external context on the text span. To this end, we replace the POS tag of head with its incoming dependency label (DL-HPB), or the combination of (the original fine-grained) POS tag and its dependency label (POS-DL-HPB). For C-HPB we use the coarse POS tag set obtained by grouping the 33 types of Chinese POS tags into 11 types following Xia (2000). For example, we generalize all verbal tags (e.g., *VA*, *VC*, *VE*, and *VV*) and all nominal tags (e.g., *NR*, *NT*, and *NN*) into *Verb* and *Noun*, respectively. We use the dependency labels in Penn2Malt which defines 9 types of dependency labels for Chinese, including *AMOD*, *DEP*, *NMOD*, *P*, *PMOD*, *ROOT*, *SBAR*, *VC*, and *VMOD*.[7]

Table 6 shows the results trained on large data. Although the number of non-terminal types decreased sharply from 63K to 3K, using the coarse POS tag set in C-HPB surprisingly lowers the performance with 1.1 BLEU scores on average (e.g., 33.75 vs. 34.86), indicating that grouping POS tags using simple linguistic rules is inappropriate for HD-HPB. We still believe that this initial negative finding should be supplemented by future work on groupping POS tags using machine learning techniques considering contextual information.

Table 6 also shows that replacing POS tags of heads with their dependency labels (DL-HPB) substantially lowers the average performance from 34.86 on BLEU score to 32.54, probably due to the very coarse granularity of the dependency labels used. In addition, replacing non-terminal label with more refined tags (e.g., combination of original POS tag and dependency label) also lowers translation performance (POS-DL-HPB). Further experiments with more fine-grained dependency labels are required.

[7] Some other types of dependency labels (e.g., *SUB*, *OBJ*) are generated from function tags which are not available in our automatic parse trees.

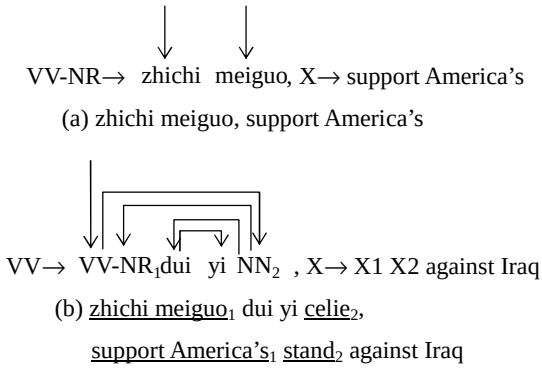

Figure 3: Examples of pharse pairs and their head-driven translation rules with dependency relation, regarding Figure 2

System	MT 03	MT 04	MT 05	MT 06	Avg.
HPB	33.59	35.39	32.20	30.60	32.95
HD-HPB	**35.50**	37.61	34.56	31.78	34.86
C-HPB	34.10	36.43	33.46	31.00	33.75
DL-HPB	32.81	35.19	32.27	29.89	32.54
POS-DL-HPB	34.08	36.78	33.14	30.43	33.61
HD-DEP-HPB	35.48	**38.17**	**34.81**	**32.38**	**35.21**

Table 6: BLEU (%) scores of models trained on large data.

5.5.3 Encoding Full Dependency Relations in Translation Rule

Xie et al. (2011) present a dependency-to-string translation model with a complete dependency structure on the source side and a moderate average improvement of 0.46 BLEU over the HPB baseline. By contrast, in our HD-HPB approach, dependency information is used to identify heads in the strings covered by non-terminals in HD-HR rules, and to refine non-terminal labels accordingly, with an average improvement of 1.91 in BLEU over the HPB baseline (when trained on the large data). This raises the question whether and to what extent complete (unlabeled) dependency information between the string and the heads in head-labeled non-terminal parts of the source side of SCFGs in HD-HPB can further improve results.

Given the source side of a translation rule (either HD-HR or NRR), say $P_s \rightarrow s_1 \dots s_m$ (where each s_i is either a terminal or a head POS in a refined non-terminal), in a further set of experiments we keep the full unlabeled dependency relations between $s_1 \dots s_m$ so as to capture contextual syntactic information in translation rules. For example, on the source side of Figure 3 (b) where *VV-NR* maps into words *zhichi* and *meiguo* while *NN* maps into word *celie*, we keep the full unlabeled dependency relations among words {*zhichi, meiguo, dui, yi, celie*}. HD-DEP-HPB (Table 6) augments translation rules in HD-HPB with full dependency relations on the source side. This further boosts the performance by 0.35 BLEU scores on average over HD-HPB and outperforms the HPB baseline by 2.26 BLEU scores on average.

5.5.4 Error Analysis

We carried out a manual error analysis comparing the outputs of our HD-HPB system with those of Chiang's (both trained on the large data). We observe that improved BLEU score often correspond to better topological ordering of phrases in the hierarchical structure of the source side, with a direct impact on which words in a source sentence should be translated first, and which later. As ungrammatical translations are often due to inappropriate topological orderings of phrases in the hierarchical structure, guiding the translation through appropriate topological ordering should improve translation quality. To give an example, consider the following input sentence from the 04 NIST MT test data and its two translation results:

- Input: 中国$_0$ 派团$_1$ 赴$_2$ 美$_3$ 采购$_4$ 二十多亿$_5$ 美元$_6$ 高$_7$ 科技$_8$ 设备$_9$
- HPB: chinese delegation to us dollar purchase of more high technology equipment
- HD-HPB: chinese delegation went to the united states to buy more us high - tech equipment

Figure 4 demonstrates the topological orderings in the two hierarchical structures. In addition to disfluency and some grammar errors (e.g., a main verb is missing), the basic HPB system also makes mistakes in reordering (e.g., 采购$_4$ 二十多亿$_5$ 美元$_6$ translated as *dollar purchase of more*). The poor translation quality, unsurprisingly, is caused by inappropriate topological ordering (Figure 4(a)). By comparison, the topological ordering reflected in the hierarchical structure of our HD-HPB model better respects syntactic structure (Figure 4(b)). Let

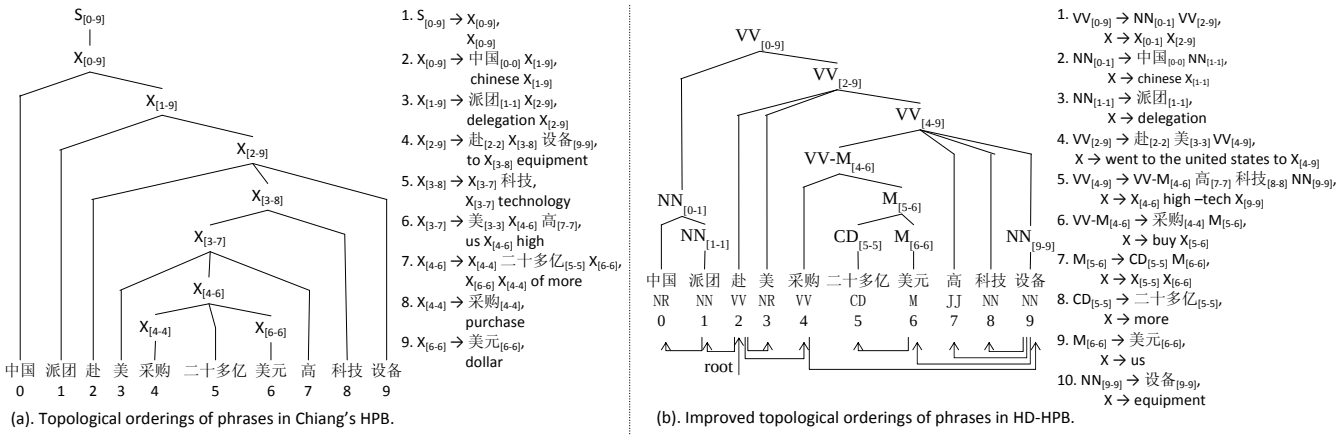

Figure 4: An example Chinese sentence and its two hierarchical structures. Note: subscript [i-j] represents spanning from word i to word j on the source side.

us refer to the HD-HPB hierarchical structure on the source side as *translation parse tree* and to the treebank-based parser derived tree as *syntactic parse tree* from which we obtain unlabeled dependency structure. Examining the translation parse trees of our HD-HPB model shows that phrases with 1/2/3/4 heads account for 64.9%/23.1%/8.8%/3.2%, respectively. Compared to 37.9% of the phrases in the translation parse trees of the HPB model, 43.2% of the phrases of our HD-HPB model correspond to a linguistically motivated constituent in the syntactic parse tree with exactly the same text span. In sum, therefore, instead of simply enforcing hard linguistic constraints imposed by a full syntactic parse structure, our model opts for a successful mix of linguistically motivated and combinatorial (matching subphrases in HPB) constraints.

6 Conclusion

In this paper, we present a head-driven hierarchical phrase-based translation model, which adopts head information (derived through unlabeled dependency analysis) in the definition of non-terminals to better differentiate among translation rules. In addition, improved and better integrated reordering rules allow better reordering between consecutive non-terminals through exploration of a larger search space in the derivation. Our model maintains the strengths of Chiang's HPB model while at the same time it addresses the over-generation problem caused by using a uniform non-terminal symbol. Experimental results on Chinese-English translation across a wide range of training and test sets demonstrate significant and consistent improvements of our HD-HPB model over Chiang's HPB model as well as over a source side version of the SAMT-style model.

Currently, we only consider head information in a word sequence. In the future work, we will exploit more syntactic and semantic information to systematically and automatically define the inventory of non-terminals (in source and target). For example, for a non-terminal symbol *VV*, we believe it will benefit translation if we use fine-grained dependency labels (subject, object etc.) used to link it to its governing head elsewhere in the translation rule.

Acknowledgments

This work was supported by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. It was also partially supported by Project 90920004 under the National Natural Science Foundation of China and Project 2012AA011102 under the "863" National High-Tech Research and Development of China. We thank the reviewers for their insightful comments.

References

Hala Almaghout, Jie Jiang, and Andy Way. 2011. CCG contextual labels in hierarchical phrase-based SMT. In *Proceedings of EAMT 2011*, pages 281–288.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of ACL 2007*, pages 33–40.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of NAACL 2000*, pages 132–139.

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.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637.

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.

Zhongjun He, Yao Meng, and Hao Yu. 2010. Maximum entropy based phrase reordering for hierarchical phrase-based translation. In *Proceedings of EMNLP 2010*, pages 555–563.

Liang Huang and David Chiang. 2005. Better k-best parsing. In *Proceedings of IWPT 2005*, pages 53–64.

Zhongqiang Huang, Martin Cmejrek, and Bowen Zhou. 2010. Soft syntactic constraints for hierarchical phrase-based translation using latent syntactic distributions. In *Proceedings of EMNLP 2010*, pages 138–147.

Philipp Koehn, Franz Josef 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.

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-HLT 2011*, pages 642–652.

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

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL 2002*, pages 295–302.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

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

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

Libin Shen, Jinxi Xu, Bing Zhang, Spyros Matsoukas, and Ralph Weischedel. 2009. Effective use of linguistic and contextual information for statistical machine translation. In *Proceedings of EMNLP 2009*, pages 72–80.

Dekai Wu. 1996. A polynomial-time algorithm for statistical machine translation. In *Proceedings of ACL 1996*, pages 152–158.

Fei Xia. 2000. The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0). Technical Report IRCS-00-07, University of Pennsylvania Institute for Research in Cognitive Science Technical.

Jun Xie, Haitao Mi, and Qun Liu. 2011. A novel dependency-to-string model for statistical machine translation. In *Proceedings of EMNLP 2011*, pages 216–226.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of NAACL 2006 - Workshop on Statistical Machine Translation*, pages 138–141.

Andreas Zollmann and Stephan Vogel. 2011. A word-class approach to labeling PSCFG rules for machine translation. In *Proceedings of ACL-HLT 2011*, pages 1–11.