Head-Driven 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

This paper presents an extension of Chiang's hierarchical phrase-based (HPB) model, called Head-Driven HPB (HD-HPB), which incorporates head information in translation rules to better capture syntax-driven information, as well as improved reordering between any two neighboring non-terminals at any stage of a derivation to explore a larger reordering search space. Experiments on Chinese-English translation on four NIST MT test sets show that the HD-HPB model significantly outperforms Chiang's model with average gains of 1.91 points absolute in BLEU.

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. 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] What is more, Chiang's HPB model suffers from limited phrase reordering combining translated 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.

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. 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 their POS to refine *X*. In addition HD-HPB provides flexible reordering rules freely mixing translation and reordering (including swap) at any stage in a derivation.

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 flex-

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

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 33–37,
Jeju, Republic of Korea, 8-14 July 2012.

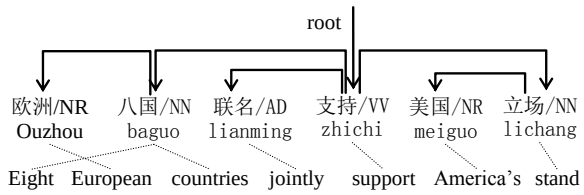

Figure 1: An example word alignment for a Chinese-English sentence pair with the dependency parse tree for the Chinese sentence. Here, each Chinese word is attached with its POS tag and Pinyin.

ible 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.

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. Instead of collapsing all non-terminals in the source language into a single symbol X as in Chiang (2007), given a word sequence f_j^i from position i to position j, we first find **heads** and then concatenate the POS tags of these heads as f_j^i's non-terminal symbol. Specifically, we adopt unlabeled 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 1 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.

According to the occurrence of terminals in translation rules, we group 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 POS tags corresponding to their heads. Given the word alignment in Figure 1, Table 1 demonstrates the difference between hierarchical rules in Chiang (2007) and HD-HRs defined here.

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 on the source side, there are four possible positional relationships for their target side translations (we use *Y* as a variable for non-terminals on the source side while all non-terminals on the target side are labeled as *X*):

- Monotone $\langle Y \rightarrow Y_1Y_2, X \rightarrow X_1X_2 \rangle$;
- Discontinuous monotone $\langle Y \rightarrow Y_1Y_2, X \rightarrow X_1 \ldots X_2 \rangle$;
- Swap $\langle Y \rightarrow Y_1Y_2, X \rightarrow X_2X_1 \rangle$;
- Discontinuous swap $\langle Y \rightarrow Y_1Y_2, X \rightarrow X_2 \ldots X_1 \rangle$.

phrase pairs	hierarchical rule	head-driven hierarchical rule
lichang, stand	X→lichang, stand	NN→lichang, X→stand
meiguo lichang$_1$, America's stand$_1$	X→meiguo X$_1$, America's X$_1$	NN→meiguo NN$_1$, X→America's X$_1$
zhichi meiguo, support America's	X→zhichi meiguo, support America's	VV-NR→zhichi meiguo, X→support America's
zhichi meiguo$_1$ lichang, support America's$_1$ stand	X→X$_1$ lichang, X$_1$ stand	VV→VV-NR$_1$ lichang, X→X$_1$ stand

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.

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.

2.3 Features and Decoding

Given e for the translation output in the target language, s and t for strings of terminals and non-terminals on the source and target side, respectively, we use a feature set analogous to the default feature set of Chiang (2007), including:

- $P_{hd\text{-}hr}(t|s)$ 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.

Our decoder is based on CKY-style chart parsing with beam search and 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's HPB model, it is possible for a non-terminal generated by a NRR to be included afterwards by a HD-HR or another NRR.

3 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 our HD-HPB and 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, 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.

We train our model on a dataset with ˜1.5M sentence pairs from the LDC dataset.[2] 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 (unlabeled) dependency structures.

We obtain the word alignments by running

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

[3]http://code.google.com/p/berkeleyparser/

[4]http://w3.msi.vxu.se/˜nivre/research/Penn2Malt.html/

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.

For evaluation, the NIST BLEU script (version 12) with the default settings is used to calculate the BLEU scores. 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.

Table 2 lists the rule table sizes. The full rule table size (including HD-HRs and NRRs) of our HD-HPB model is ˜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 ˜0.7 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 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 1.91 in BLEU (and similar improvements over Moses HPB).

Table 3 shows that the head-driven scheme outperforms a SAMT-style approach (for each test set $p < 0.01$), indicating that head information is more effective than (partial) CFG categories. Taking *lianming zhichi* in Figure 1 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*[5] which is more susceptible to data sparsity. 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*.

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

System	Total	MT 03	MT 04	MT 05	MT 06	Avg.
HPB	39.6	2.8	4.7	3.3	3.0	3.4
HD-HPB	59.5/0.6	1.9/0.1	3.4/0.2	2.3/0.2	2.0/0.1	2.4/0.2

Table 2: Rule table sizes (in million) of different models. Note: 1) For HD-HPB, the rule sizes separated by / indicate HD-HRs and NRRs, respectively; 2) Except for "Total", the figures correspond to rules filtered on the corresponding test set.

System	MT 03	MT 04	MT 05	MT 06	Avg.
Moses HPB	32.94*	35.16	32.18	29.88*	32.54
HPB	33.59	35.39	32.20	30.60	32.95
HD-HPB	**35.50****	**37.61****	**34.56****	**31.78****	**34.86**
SAMT-HPB	34.07	36.52**	32.90*	30.66	33.54
HD-HR+Glue	34.58**	36.55**	33.84**	31.06	34.01

Table 3: BLEU (%) scores of different models. Note: 1) SAMT-HPB indicates our HD-HPB model with non-terminal scheme of Zollmann and Venugopal (2006); 2) HD-HR+Glue indicates our HD-HPB model replacing NRRs with glue rules; 3) Significance tests for Moses HPB, HD-HPB, SAMT-HPB, and HD-HR+Glue are done against HPB.

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). Table 3 shows that on average more than half of the improvement over HPB (Chiang and Moses) comes from the refined HD-HRs, the rest from NRRs.

Examining translation rules extracted from the training data shows that there are 72,366 types of non-terminals with respect to 33 types of POS tags. 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.

4 Conclusion

We present a head-driven hierarchical phrase-based (HD-HPB) translation model, which adopts head information (derived through unlabeled dependency analysis) in the definition of non-terminals to better differentiate among translation rules. In ad-

dition, improved and better integrated reordering rules allow better reordering between consecutive non-terminals through exploration of a larger search space in the derivation. Experimental results on Chinese-English translation across four test sets demonstrate significant improvements of the HD-HPB model over both Chiang's HPB and a source-side SAMT-style refined version of HPB.

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.

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. 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.

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.