Lexicalized Reordering for Left-to-Right Hierarchical Phrase-based Translation

Maryam Siahbani[*]
Department of Computer Information Systems
University of the Fraser Valley
Abbotsford BC, Canada
maryam.siahbani@ufv.ca

Anoop Sarkar
School of Computing Science
Simon Fraser University
Burnaby BC, Canada
anoop@cs.sfu.ca

Abstract

Phrase-based and hierarchical phrase-based (Hiero) translation models differ radically in the way reordering is modeled. Lexicalized reordering models play an important role in phrase-based MT and such models have been added to CKY-based decoders for Hiero. Watanabe et al. (2006) propose a promising decoding algorithm for Hiero (LR-Hiero) that visits input spans in arbitrary order and produces the translation in left to right (LR) order which leads to far fewer language model calls and leads to a considerable speedup in decoding. We introduce a novel shift-reduce algorithm to LR-Hiero to decode with our lexicalized reordering model (LRM) and show that it improves translation quality for Czech-English, Chinese-English and German-English.

1 Introduction

Phrase-based machine translation handles reordering between source and target languages by visiting phrases in the source in arbitrary order while generating the target from left to right. A distortion penalty is used to penalize deviation from the monotone translation (no reordering) (Koehn et al., 2003; Och and Ney, 2004). Identical distortion penalties for different types of phrases ignore the fact that certain phrases (with certain words) were more likely to reorder than others. State-of-the-art phrase based translation systems address this issue by applying a *lexicalized reordering model* (LRM) (Tillmann, 2004; Koehn et al., 2007; Galley and Manning, 2008; Galley and Manning, 2010) which uses word aligned data to score phrase pair reordering. These models distinguish three orientations with respect to the previously translated phrase: *monotone* (M), *swap* (S), and *discontinuous* (D), which are primarily designed to handle local re-orderings of neighbouring phrases.

Hierarchical phrase-based translation (Hiero) (Chiang, 2007) uses hierarchical phrases for translations represented as lexicalized synchronous context-free grammar (SCFG). Non-terminals in the SCFG rules correspond to gaps in phrases which are recursively filled by other rules (phrases). The SCFG rules are extracted from word and phrase alignments of a bitext. Hiero uses CKY-style decoding which parses the source sentence with time complexity $O(n^3)$ and synchronously generates the target sentence (translation).

Watanabe et al. (2006) proposed a left-to-right (LR) decoding algorithm for Hiero (LR-Hiero) which follows the Earley (Earley, 1970) algorithm to parse the source sentence and synchronously generate the translation in a left-to-right manner. This algorithm is combined with beam search and has time complexity $O(n^2 b)$ where n is the length of source sentence and b is the size of beam (Huang and Mi, 2010). LR-Hiero constrains the SCFG rules to be prefix-lexicalized on the target side aka Greibach Normal Form (GNF). Throughout this paper we abuse the notation for simplicity and use the term GNF grammars for such SCFGs. This leads to a single language model (LM) history for each hypothesis and speeds up decoding significantly, up to four times faster (Siahbani et al., 2013).

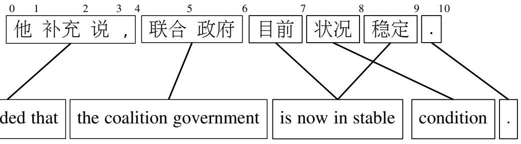
The Hiero translation model handles reordering very differently from a phrase-based model, through weighted translation rules (SCFGs) determined by non-terminal mappings. The rule $X \rightarrow \langle ne\ X_1\ pas, do\ not\ X_1 \rangle$ indicates the translation of the phrase between *ne* and *pas* will be after the English phrase *do not*. However, reordering features can also be added to the Hiero log-linear translation model. Siahbani et al. (2013) introduce a new distortion feature to Hiero and LR-Hiero which

[*] This work was done while the first author was a Ph.D. student at SFU.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 612–618,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

rules	hypotheses ⟨h_t, h_s, h_c⟩
	⟨<s>,⟦[0,10]⟧,0⟩
1) X→⟨他 补充 说 ，X_1/ *He added that* X_1⟩	⟨<s> *He added that*,⟦[4,10]⟧,4.3⟩
2) X→⟨联合 政府 X_1/ *the coalition government* X_1⟩	⟨<s> *He added that the coalition government*,⟦[6,10]⟧,7.7⟩
3) X→⟨目前 X_1稳定 X_2/ *is now in stable* $X_1 X_2$⟩	⟨<s> *He added that the coalition government is now in stable*,⟦[7,8][9,10]⟧,11.2⟩
4) X→⟨状况 / *condition*⟩	⟨<s> *He added that the coalition government is now in stable condition*,⟦[9,10]⟧,13.4⟩
5) X→⟨./.⟩	⟨<s> *He added that the coalition government is now in stable condition*.</s> ,⟦⟧,14.3⟩

Figure 1: The process of translating a Chinese (Fig. 2) sentence to English using LR-Hiero. Left side shows the rule used in each step of creating the derivation. The hypotheses column shows 3-tuple partial hypotheses: the translation prefix, h_t, the ordered list of yet-to-be-covered spans, h_s, and cost h_c.



1)⟨他 补充 说 ，/ *He added that*⟩
2)⟨联合 政府 / *the coalition government*⟩
3)⟨目前 *_* 稳定 / *is now in stable*⟩
4)⟨状况 / *condition*⟩
5)⟨./.⟩

Figure 2: A word-aligned Chinese-English sentence pair on the top (from devset data used in experiments.) The source-target phrase pairs created by removing the non-terminals from the rules used in decoding (Fig. 1) are shown on the bottom.

significantly improves translation quality in LR-Hiero and improves Hiero results to a lesser extent. Nguyen and Vogel (2013) integrate phrase-based distortion and lexicalized reordering features with CKY-based Hiero decoder which significantly improve the translation quality. In their approach, each partial hypothesis during decoding is mapped into a sequence of phrase-pairs then the distortion and reordering features are computed similar to phrase-based MT. They use a LRM trained for phrase-based MT (Galley and Manning, 2010) which applies some restrictions on the Hiero rules. (Cao et al., 2014; Huck et al., 2013) propose different approaches to directly train LRM for Hiero rules. However, these approaches are designed for CKY-decoding and cannot be directly used or adapted for LR-Hiero decoding which uses an Earley-style parsing algorithm. The crucial difference is the nature of bottom-up versus left to right decisions for lexicalized reordering and generating the translation in left-to-right manner. In this paper, we introduce a novel shift-reduce algorithm to learn a lexicalized reordering model (LRM) for LR-Hiero. We show that augmenting LR-Hiero with an LRM improves translation quality for Czech-English, significantly improves results for Chinese-English and German-English, while performing three times fewer language model queries on average, compared to CKY-Hiero.

2 Lexicalized Reordering for LR-Hiero

The main idea in phrase-based LRM is to divide possible reorderings into three orientations that can be easily determined during decoding and also from word-aligned sentence pairs (parallel corpus). Given a source sentence **f**, a sequence of target language phrases **e** $= (\bar{e_1}, \ldots, \bar{e_n})$ is generated by the decoder. A phrase alignment **a**$= (a_1, \ldots a_n)$ defines a source phrase $\bar{f}_{a_i}$ for each target phrase $\bar{e}_i$. For each phrase-pair $\langle \bar{f}_{a_i}, e_i \rangle$, the orientations are described in terms of the previously translated source phrase $\bar{f}_{a_{i-1}}$:
Monotone (M): $\bar{f}_{a_i}$ immediately follows $\bar{f}_{a_{i-1}}$.
Swap (S): $\bar{f}_{a_{i-1}}$ immediately follows $\bar{f}_{a_i}$.
Discontinuous (D): $\bar{f}_{a_i}$ and $\bar{f}_{a_{i-1}}$ are not adjacent in the source sentence.
We only define the left-to-right case here; the right-to-left case ($\bar{f}_{a_{i+1}}$) is symmetrical. The probability of an orientation given a phrase pair $\langle \bar{f}, \bar{e} \rangle$ can be estimated using relative frequency:

$$P(o|\bar{f}, \bar{e}) = \frac{cnt(o, \bar{f}, \bar{e})}{\sum_{o' \in \{M,S,D\}} cnt(o', \bar{f}, \bar{e})} \qquad (1)$$

where, $o \in \{M, S, D\}$ and *cnt* is computed on word-aligned parallel data (count phrase-pairs and their orientations). Given the sparsity of the orientation types, we use smoothing. As the decoder develops a new hypothesis by translating a source phrase, $\bar{f}_{a_i}$, it scores the orientation, o_i wrt a_{i-1}. The log probability of the orientation is added as a feature function to the log-linear translation model.

LR-Hiero uses a subset of the Hiero SCFG rules where the target rules are in Greibach Normal Form (GNF): $\langle \gamma, \bar{e}\,\beta \rangle$ where γ is a string of non-terminal and source words, $\bar{e}$ is a target phrase and β is a possibly empty sequence of non-terminals. We abuse notation slightly and call this a GNF SCFG grammar. In LR-Hiero each hypothesis consists of a translation prefix, h_t, an ordered sequence of untranslated spans on the source sen-

rules	$r_i.\bar{f}$	O_i	S
	$\{-1\}$		[(-1)-(-1)]
1) $\langle$0 1 2 3 4 X_1/ under such circumstance $X_1\rangle$	$\{0, 1, 2, 3, 4\}$	M	[(-1)-4]
2) $\langle$5 X_1/, $X_1\rangle$	$\{5\}$	M	[(-1)-5]
3) $\langle$6 X_1 11/when $X_1\rangle$	$\{6, 11\}$	M	[(-1)-11]
4) $\langle$7 8 X_1/the right of life $X_1\rangle$	$\{7, 8\}$	D	[(-1)-11]
5) $\langle$9 10/was deprived$\rangle$	$\{9, 10\}$	M	[(-1)-11]
6) $\langle$12 X_1/, $X_1\rangle$	$\{12\}$	M	[(-1)-12]
7) $\langle$13 14 X_1/it can only $X_1\rangle$	$\{13, 14\}$	M	[(-1)-14]
8) $\langle$15 16$X_1$18/take violence$X_1\rangle$	$\{15, 16, 18\}$	M	[(-1)-18]
9) $\langle$17/to$\rangle$	$\{17\}$	D	[(-1)-18]

Figure 3: Computing correct orientation for each rule during decoding in LR-Hiero for the example in Fig. 4. **rules**: the rules used in the derivation. $r_i.\bar{f}$: the position of rule's lexical terms in the source sentence; O_i: the identified orientation. S is the recent translated source span (possibly discontinuous). At each step O_i is identified by comparing $r_i.\bar{f}$ to S in the previous step or last translated source phrase $r_{i-1}.\bar{f}$.

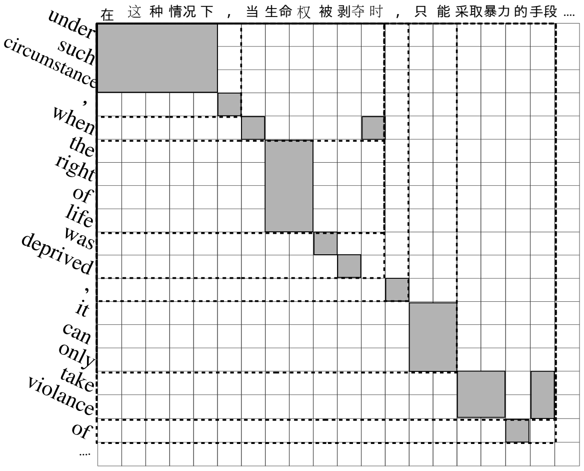


Figure 4: An example showing that the shift-reduce algorithm can capture local reorderings like: *the right of life* and *was deprived.*

tence, h_s and a numeric cost, h_c. The initial hypothesis consists of an empty translation ($\langle s \rangle$), a span of the whole source sentence and cost 0 (Figure 1). To develop a new hypothesis from a current hypothesis, the LR-Hiero decoder applies a GNF rule to the first untranslated span, $h_s[0]$, of old hypothesis. The translation prefix of the new hypothesis is generated by appending the target side of the applied rule, $\bar{e}$, to the translation prefix of the old hypothesis, h_t. Corresponding to the applied rule, the uncovered spans of the old hypothesis are also updated and assigned to the new hypothesis (Figure 1).

Target generation in LR-Hiero is analogous to phrase-based MT. Given an input sentence **f**, the output translation is a sequence of contiguous target-language phrases $\mathbf{e} = (\bar{e_1}, \ldots, \bar{e_n})$ incrementally concatenated during decoding. We can define a phrase alignment $\mathbf{a} = (a_1, \ldots a_n)$ which align each target phrase, $\bar{e_i}$ to a source phrase f_{a_i} corresponding to source side of a rule, r_i used at step i. But unlike target, source phrases can be discontiguous. Figure 1 illustrates the process of translating a Chinese-English sentence pair by LR-Hiero. Corresponding to each rule a phrase pair can be created (shown in Figure 2). The final translation is the ordered sequence of target side of these phrase pairs. Although the target generation is similar to phrase-based MT, the LR-Hiero decoder parse the source sentence using the SCFG rules and the order for translating source spans is determined by the grammar. However the LR-Hiero decoder uses an Earley-style parsing algorithm and unlike CKY does not utilise translated smaller spans to generate translations for bigger spans bottom-up.

2.1 Training

We compute $P(o|\bar{f}, \bar{e})$, which is the probability of an orientation given phrase pair of a rule, $r.p = \langle \bar{f}, \bar{e} \rangle$, on word-aligned data using relative frequency. We assume that phrase $\bar{e}$ spans the word range $s \ldots t$ in the target sentence and the phrase $\bar{f}$ spans the range $u \ldots v$ in the source sentence.

For a given phrase pair $\langle \bar{f}, \bar{e} \rangle$, we set $o = M$ if there is a phrase pair,$\langle \bar{f'}, \bar{e'} \rangle$, where its target side, $\bar{e'}$, appears just before the target side of the given phrase, $\bar{e}$, or $s = t' + 1$, and its source side, $\bar{f'}$, also appears just before $\bar{f}$, or $u = v' + 1$. Orientation is S if there is a phrase pair, $\langle \bar{f'}, \bar{e'} \rangle$, where $\bar{e'}$ appears just before $\bar{e}$, or $s = t' + 1$, and $\bar{f'}$ appears just after $\bar{f}$, or $v = u' - 1$. Otherwise orientation is D. We consider phrase pairs of any length to compute orientation. Note that although phrase pairs extracted from the rules that can be discontinuous (on source), just continuous source phrases in each sentence pair are used to compute orientation (previously translated phrases). Once orientation counts for rules (phrase-pairs obtained form rules) are collected from the bitext, the probability model $P(o|\bar{f}, \bar{e})$ is estimated using recursive MAP smoothing as discussed in (Cherry, 2013).

2.2 Decoding

Phrase-based LRM uses local information to determine orientation for a new phrase pair, $\langle \bar{f}_{a_i}, \bar{e_i} \rangle$, during decoding (Koehn et al., 2007; Tillmann, 2004). For left-to-right order, $\bar{f}_{a_i}$ is compared to the previously translated phrase $\bar{f}_{a_{i-1}}$. Galley and Manning (2008) introduce the hierarchical phrase reordering model (HRM) which increases the consistency of orientation assignments. In HRM, the emphasis on the previously translated phrase is removed and instead a compact representation of the full translation history, as represent by a shift-reduce stack, is used. Once a source span is translated, it is shifted onto the stack; if the two spans on the top are adjacent, then a reduction merges the two. During decoding, orientations are always determined with respect to the top of this stack, rather than the previously translated phrase.

Although we reduce rules to phrase pairs to train the reordering model, LR-Hiero decoder uses SCFG rules for translation and the order of source phrases (spans) are determined by the non-terminals in SCFG rules. Therefore we cannot simply rely on the previously translated phrase to compute the orientation and reordering scores. Since LR-Hiero uses lexicalized glue rules (Watanabe et al., 2006), non-terminals can be matched to very long spans on the source sentence. It makes LRM in LR-Hiero comparable to HRM in phrase-based MT. However, we cannot rely on the full translation history like HRM, since translation model is a SCFG grammar encoding reordering information.

We employ a shift-reduce approach to find a compact representation of the recent translated source spans which is also represented by a stack, S, for each hypothesis. However, S always contains just one source span (which might be discontiguous), unlike HRM which maintains all previously translated solid spans (In Figure 4, the dotted lines shows the only span in the stack during LR-Hiero decoding). As the decoder applies a rule, r_i, the corresponding source phrase $r_i.\bar{f}$ is compared respect to the span in S to determine the orientation. If they are adjacent or S covers the span $r_i.\bar{f}$, they are reduced. Otherwise stack is set to the span of new rule, $S = r_i.\bar{f}$. The orientation of $r_i.\bar{f}$ is computed with respect to S but if they are not adjacent (*M* or *S*), we still need to consider the possible local reordering with respect to the previous rule $r_{i-1}.\bar{f}$. In Figure 3, rules #5,#4 are monotone, while both are covered by the current span in S. Since the stack always contains one span, this algorithm runs in $O(1)$. Therefore, only a limited number of comparisons is used to update S and compute orientation. Unlike HRM which needs to maintain a sequence of contiguous spans in the stack and runs in linear time.

Figure 3 illustrates the application of shift-reduce approach to compute orientation for initial decoding steps of a Chinese-English sentence pair shown in Figure 4. We show source words in the rules with the corresponding index in the source sentence. S and $r_i.\bar{f}$ for the initial hypothesis are set to -1, corresponding to the start of sentence symbol, making it easy to compute the correct orientation for spans at the beginning of the input (with index 0).

3 Experiments

We evaluate lexicalized reordering model for LR-Hiero on three language pairs: German-English (De-En), Czech-English (Cs-En) and Chinese-English (Zh-En). Table 1 shows the corpus statistics for all language.

We train a 5-gram LM on the Gigaword corpus using KenLM (Heafield, 2011). The weights in the log-linear model are tuned by minimizing BLEU loss through MERT (Och, 2003) on the dev set for each language pair and then report BLEU scores on the test set. Pop limit for Hiero and LR-Hiero is 500 and beam size for Moses is 1000. Other extraction and decoder settings such as maximum phrase length, etc. are identical across different settings.

We use 3 baselines in our experiments:

- **Hiero:** we use our in-house implementation of Hiero, *Kriya*, in Python (Sankaran et al., 2012). Kriya can obtain statistically significantly equal BLEU scores when compared with Moses (Koehn et al., 2007) for several language pairs (Razmara et al., 2012; Callison-Burch et al., 2012).

- **phrase-based:** Moses (Koehn et al., 2007) with and without lexicalized reordering features.

- **LR-Hiero:** LR-Hiero decoding with cube pruning and queue diversity of 10 (Siahbani and Sarkar, 2014b).

To make the results comparable we use the standard SMT features for log-linear model in translation systems. relative-frequency translation probabilities $p(f|e)$ and $p(e|f)$, lexical translation probabilities $p_l(f|e)$ and $p_l(e|f)$, a language model probability, word count, phrase count and distortion. In addition, two distortion features proposed

	Corpus	Train/Dev/Test
Cs-En	Europarl.v7; CzEng.v0.9; News commentary(nc) 2008,2009,2011	7.95M/3000/3003
De-En	Europarl.v7; WMT2006	1.5M/2000/2000
Zh-En	HK + GALE ph1; MTC 1,3,4	2.3M/1928/919

Table 1: Corpus statistics in number of sentences. Tuning and test sets for Chinese-English has 4 references.

Model	Cs-En	De-En	Zh-En
Hiero	6279.3	7152.3	6524.7
LR-Hiero + LRM	**2015.1**	**2908.3**	**2225.7**

Table 2: Translation time in terms of average number of LM queries.

Model	Cs-En	De-En	Zh-En
Phrase-based	20.32	24.71	25.68
+ LRM	20.74	**25.99**	26.61
Hiero	20.77	25.72	**27.65**
LR-Hiero	20.52	24.96	25.73
+ NVLRM	20.49	24.98	25.9
+ LRM	**20.86**	25.44	26.57

Table 3: Translation accuracy in terms of BLEU for different baselines and LR-Hiero with lexicalized reordering model. The rows are grouped such that each group use the same model.

by (Siahbani et al., 2013) are added to both Hiero and LR-Hiero. The LRM proposed in this paper uses a GNF grammar and LR decoding, therefore we apply it only to LR-Hiero. The GNF rules are obtained from word and phrase aligned bitext using the rule extraction algorithm proposed by (Siahbani and Sarkar, 2014a).

Table 3 compares the performance of different translation systems in terms of translation quality (BLEU). In all language pairs the proposed lexicalized reordering model improves the translation quality of LR-Hiero. These observations are comparable to the effect of LRM in phrase-based translation system. In Cs-En, LRM gets the best results and it significantly improves the the LR-Hiero results for De-En and Zh-En (p-value$<$0.05, evaluated by MultEval (Clark et al., 2011)). To compare our approach to Nguyen and Vogel (2013), we adopt their algorithm to LR-Hiero and use the same LRM trained for GNF rules (marked as *NVLRM* in Table 3). Unsurprisingly this approach could not improve the translation quality in LR-Hiero. This approach computes the LRM for all candidate translation of each span after obtaining the full translations. In bottom-up decoders it helps to prune the hypotheses effectively while in LR-Hiero decoder as we apply a rule before knowing the translation of smaller spans the computation of LRM will be postponed and gets less effective in decoding.

Table 2 shows the performance in terms of decoding speed. We use the same wrapper for Hiero and LR-Hiero to query the language model and report the average on a sample set of 50 sentences from test sets. We can see LR-Hiero+LRM still works 3 times faster than Hiero in terms of number of LM calls which leads to a faster decoder speed.

4 Conclusion

We have proposed a novel lexicalized reordering model (LRM) for the left-to-right variant of Hiero called LR-Hiero distinct from previous LRM models. The previous LRM models proposed for Hiero are just applicable to bottom-up decoders like CKY. We proposed a model for the left-to-right decoding algorithm of LR-Hiero. We showed that our novel shift-reduce algorithm to decode with the lexicalized reordering model significantly improved the translation quality of LR-Hiero on three different language pairs.

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. The research was also partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN 262313 and RGPAS 446348) to the second author.

References

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June. Association for Computational Linguistics.

Hailong Cao, Dongdong Zhang, Mu Li, Ming Zhou, and Tiejun Zhao. 2014. A lexicalized reordering model for hierarchical phrase-based translation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1144–1153, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Colin Cherry. 2013. Improved reordering for phrase-based translation using sparse features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 22–31, Atlanta, Georgia, June. Association for Computational Linguistics.

David Chiang. 2007. Hierarchical phrase-based translation. *Comput. Linguist.*, 33(2):201–228, June.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA, June. Association for Computational Linguistics.

Jay Earley. 1970. An efficient context-free parsing algorithm. *Commun. ACM*, 13(2):94–102, February.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2008, pages 848–856, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michel Galley and Christopher D. Manning. 2010. Accurate non-hierarchical phrase-based translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 966–974, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 187–197, Stroudsburg, PA, USA. Association for Computational Linguistics.

Liang Huang and Haitao Mi. 2010. Efficient incremental decoding for tree-to-string translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 273–283, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthias Huck, Joern Wuebker, Felix Rietig, and Hermann Ney. 2013. A phrase orientation model for hierarchical machine translation. In *ACL 2013 Eighth Workshop on Statistical Machine Translation*, pages 452–463, Sofia, Bulgaria, August.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

ThuyLinh Nguyen and Stephan Vogel. 2013. Integrating phrase-based reordering features into a chart-based decoder for machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1587–1596, Sofia, Bulgaria, August. Association for Computational Linguistics.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Majid Razmara, Baskaran Sankaran, Ann Clifton, and Anoop Sarkar. 2012. Kriya - the sfu system for translation task at wmt-12. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 356–361, Montréal, Canada, June. Association for Computational Linguistics.

Baskaran Sankaran, Majid Razmara, and Anoop Sarkar. 2012. Kriya - an end-to-end hierarchical phrase-based MT system. *Prague Bull. Math. Linguistics*, 97:83–98.

Maryam Siahbani and Anoop Sarkar. 2014a. Expressive hierarchical rule extraction for left-to-right translation. In *Proceedings of the Eleventh Conference of the Association for Machine Translation in the Americas (AMTA)*, volume 1, pages 1–14.

Maryam Siahbani and Anoop Sarkar. 2014b. Two improvements to left-to-right decoding for hierarchical phrase-based machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 221–226, Doha, Qatar, October. Association for Computational Linguistics.

Maryam Siahbani, Baskaran Sankaran, and Anoop Sarkar. 2013. Efficient left-to-right hierarchical phrase-based translation with improved reordering. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1089–1099, Seattle, Washington, USA, October. Association for Computational Linguistics.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 101–104, Stroudsburg, PA, USA. Association for Computational Linguistics.

Taro Watanabe, Hajime Tsukada, and Hideki Isozaki. 2006. Left-to-right target generation for hierarchical phrase-based translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 777–784, Sydney, Australia, July. Association for Computational Linguistics.