Improved Reordering for Shallow-*n* Grammar based Hierarchical Phrase-based Translation

Baskaran Sankaran and Anoop Sarkar
School of Computing Science
Simon Fraser University
Burnaby BC. Canada
{baskaran, anoop}@cs.sfu.ca

Abstract

Shallow-n grammars (de Gispert et al., 2010) were introduced to reduce over-generation in the Hiero translation model (Chiang, 2005) resulting in much faster decoding and restricting reordering to a desired level for specific language pairs. However, Shallow-n grammars require parameters which cannot be directly optimized using minimum error-rate tuning by the decoder. This paper introduces some novel improvements to the translation model for Shallow-n grammars. We introduce two rules: a BITG-style reordering *glue* rule and a simpler monotonic concatenation rule. We use separate features for the new rules in our log-linear model allowing the decoder to directly optimize the feature weights. We show this formulation of Shallow-n hierarchical phrase-based translation is comparable in translation quality to full Hiero-style decoding (without shallow rules) while at the same time being considerably faster.

1 Introduction

Hierarchical phrase-based translation (Chiang, 2005; Chiang, 2007) extends the highly lexicalized models from phrase-based translation systems in order to model lexicalized reordering and discontiguous phrases. However, a major drawback in this approach, when compared to phrase-based systems, is the total number of rules that are learnt are several orders of magnitude larger than standard phrase tables, which leads to over-generation and search errors and contribute to much longer decoding times. Several approaches have been proposed to address these issues: from filtering the extracted synchronous grammar (Zollmann et al., 2008; He et al., 2009; Iglesias et al., 2009) to alternative Bayesian approaches for learning minimal grammars (Blunsom et al., 2008; Blunsom et al., 2009; Sankaran et al., 2011). The idea of Shallow-n grammars (de Gispert et al., 2010) takes an orthogonal direction for controlling the over-generation and search space in Hiero decoder by restricting the degree of nesting allowed for Hierarchical rules.

We propose an novel statistical model for Shallow-*n* grammars which does not require additional non-terminals for monotonic re-ordering and also eliminates hand-tuned parameters and instead introduces an automatically tunable alternative. We introduce a BITG-style (Saers et al., 2009) reordering glue rule (§ 3) and a monotonic X-glue rule (§ 4). Our experiments show the resulting Shallow-n decoding is comparable in translation quality to full Hiero-style decoding while at the same time being considerably faster.

All the experiments in this paper were done using *Kriya* (Sankaran et al., 2012) hierarchical phrase-based system which also supports decoding with Shallow-n grammars. We extended Kriya to additionally support reordering glue rules as well.

2 Shallow-*n* Grammars

Formally a Shallow-*n* grammar G is defined as a 5-tuple: $G = (N, T, R, R_g, S)$, such that T is a set of finite terminals and N a set of finite non-terminals $\{X^0, \ldots, X^N\}$. R_g refers to the glue rules that rewrite the start symbol S:

$$S \rightarrow <X,\ X> \quad (1)$$

$$S \rightarrow <SX,\ SX> \quad (2)$$

R is the set of finite production rules in G and has two types, viz. hierarchical (3) and terminal (4). The hierarchical rules at each level n are additionally conditioned to have *at least* one X^{n-1} non-terminal

in them. $\sim$ represents the indices for aligning non-terminals where co-indexed non-terminal pairs are rewritten synchronously.

$$X^n \rightarrow <\gamma,\ \alpha,\ \sim>,\ \gamma, \alpha \in \{\{X^{n-1}\} \cup T^+\} \quad (3)$$

$$X^0 \rightarrow <\gamma,\ \alpha>, \qquad \gamma, \alpha \in T^+ \quad (4)$$

de Gispert et al. (2010) also proposed additional non-terminals M^k to enable reordering over longer spans by concatenating the hierarchical rules within the span. It also uses additional parameters such as monotonicity level (K_1 and K_2), maximum and minimum rule spans allowed for the non-terminals (§3.1 and 3.2 in de Gispert et al. (2010)). The monotonicity level parameters determine the number of non-terminals that are combined in monotonic order at the $N-1$ level and can be adapted to the reordering requirements of specific language pairs. The maximum and minimum rule spans further control the usage of hierarchical rule in a derivation by stipulating the underlying span to be within a range of values. Intuitively, this avoids hierarchical rules being used for a source phrase that is either too short or too long. While these parameters offer flexibility for adapting the translation system to specific language pairs, they have to be manually tuned which is tedious and error-prone.

We propose an elegant and automatically tunable alternative for the Shallow-*n* grammars setting. Specifically, we introduce a BITG-style reordering glue rule (§ 3) and a monotonic X-glue rule (§ 4). Our experiments show the resulting Shallow-n decoding to perform to the same level as full-Hiero decoding at the same time being faster.

In addition, our implementation of Shallow-n grammar differs from (de Gispert et al., 2010) in at least two other aspects. First, their formulation constrains the X in the glue rules to be at the top-level and specifically they define them to be: $S \rightarrow <SX^N, SX^N>$ and $S \rightarrow <X^N, X^N>$, where X^N is the non-terminal corresponding to the top-most level. Interestingly, this resulted in poor BLEU scores and we found the more generic glue rules (as in (1) and (2)) to perform significantly better, as we show later.

Secondly, they also employ pattern-based filtering (Iglesias et al., 2009) in order to reducing redundancies in the Hiero grammar by filtering it based on certain rule patterns. However in our limited experiments, we observed the filtered grammar to perform worse than the full grammar, as also noted by (Zollmann et al., 2008). Hence, we do not employ any grammar filtering in our experiments.

3 Reordering Glue Rule

In this paper, we propose an additional BITG-style glue rule (called R-glue) as in (5) for reordering the phrases along the left-branch of the derivation.

$$S \rightarrow <SX,\ XS> \quad (5)$$

In order to use this rule sparsely in the derivation, we use a separate feature for this rule and apply a penalty of 1. Similar to the case of regular glue rules, we experimented with a variant of the reordering glue rule, where X is restricted to the top-level: $S \rightarrow <SX^N, X^N S>$ and $S \rightarrow <X^N, X^N>$.

3.1 Language Model Integration

The traditional phrase-based decoders using beam search generate the target hypotheses in the left-to-right order. In contrast, Hiero-style systems typically use CKY chart-parsing decoders which can freely combine target hypotheses generated in intermediate cells with hierarchical rules in the higher cells. Thus the generation of the target hypotheses are fragmented and out of order compared to the left to right order preferred by n-gram language models.

This leads to challenges in the estimation of language model scores for partial target hypothesis, which is being addressed in different ways in the existing Hiero-style systems. Some systems add a sentence initial marker (<s>) to the beginning of each path and some other systems have this implicitly in the derivation through the translation models. Thus the language model scores for the hypothesis in the intermediate cell are approximated, with the true language model score (taking into account sentence boundaries) being computed in the last cell that spans the entire source sentence.

We introduce a novel improvement in computing the language model scores: for each of the target hypothesis fragment, our approach finds the best position for the fragment in the final sentence and uses the corresponding score. We compute three different scores corresponding to the three positions where the fragment can end up in the final sentence, viz.

sentence initial, middle and final: and choose the best score. As an example for fragment t_f consisting of a sequence of target tokens, we compute LM scores for i) <s> t_f, ii) t_f and iii) t_f </s> and use the best score for pruning alone[1].

This improvement significantly reduces the search errors while performing *cube pruning* (Chiang, 2007) at the cost of additional language model queries. While this approach works well for the usual glue rules, it is particularly effective in the case of reordering glue rules. For example, a partial candidate covering a non-final source span might translate to the final position in the target sentence. If we just compute the LM score for the target fragment as is done normally, this might get pruned early on before being reordered by the new glue rule. Our approach instead computes the three LM scores and it would correctly use the last LM score which is likely to be the best, for pruning.

4 Monotonic Concatenation Glue rule

The reordering glue rule facilitates reordering at the top-level. However, this is still not sufficient to allow long-distance reordering as the shallow-decoding restricts the depth of the derivation. Consider the Chinese example in Table 1, in which translation of the Chinese word corresponding to the English phrase *the delegates* involves a long distance reordering to the beginning of the sentence. Note that, three of the four human references prefer this long distance reordering, while the fourth one avoids the movement by using a complex construction with relative clause and a sentence initial prepositional phrase.

Such long distance reordering is very difficult in conventional Hiero decoding and more so with the Shallow-n grammars. While the R-glue rule permit such long distance movements, it also requires a long phrase generated by a series of rules to be moved as a block. We address this issue, by adding a monotonic concatenation (called X-glue) rule that concatenates a series of hierarchical rules. In order to control overgeneration, we apply this rule only at the $N-1$ level similar to de Gispert et al. (2010).

$$X^{N-1} \rightarrow <X^{N-1}X^{N-1},\ X^{N-1}X^{N-1}> \quad (6)$$

[1]This ensures the the LM score estimates are never underestimated for pruning. We retain the LM score for fragment (case ii) for estimating the score for the full candidate sentence later.

However unlike their approach, we use this rule as a feature in the log-linear model so that its weight can be optimized in the tuning step. Also, our approach removes the need for additional parameters K_1 and K_2 for controlling monotonicity, which was being tuned manually in their work. For the Chinese example above, shallow-1 decoding using R and X-glue rules achieve the complex movement resulting in a significantly better translation than full-Hiero decoding as shown in the last two lines in Table 1.

5 Experiments

We present results for Chinese-English translation as it often requires heavy reordering. We use the HK parallel text and GALE phase-1 corpus consisting of ∼2.3M sentence pairs for training. For tuning and testing, we use the MTC parts 1 and 3 (1928 sentences) and MTC part 4 (919 sentences) respectively. We used the usual pre-processing pipeline and an additional segmentation step for the Chinese side of the bitext using the LDC segmenter[2].

Our log-linear model uses the standard features conditional ($p(e|f)$ and $p(f|e)$) and lexical ($p_l(e|f)$ and $p_l(f|e)$) probabilities, phrase (p_p) and word (w_p) penalties, language model and regular glue penalty (m_g) apart from two additional features for $R-$glue (r_g) and $X-$glue (x_g).

Table 2 shows the BLEU scores and decoding time for the MTC test-set. We provide the IBM BLEU (Papineni et al., 2002) scores for the Shallow-n grammars for order: $n = 1, 2, 3$ and compare it to the full-Hiero baseline. Finally, we experiment with two variants of the S glue rules, i) a restricted version where the glue rules combine only X at level N, (column 'Glue: X^N' in table), ii) more free variant where they are allowed to use any X freely (column 'Glue: X' in table).

As it can be seen, the unrestricted glue rules variant (column 'Glue: X') consistently outperforms the glue rules restricted to the top-level non-terminal X^N, achieving a maximum BLEU score of 26.24, which is about 1.4 BLEU points higher than the latter and is also marginally higher than full Hiero. The decoding speeds for free-Glue and restricted-Glue variants were mostly identical and so we only provide the decoding time for the latter. Shallow-2 and

[2]We slightly modified the LDC segmenter, in order to correctly handle non-Chinese characters in ASCII and UTF8.

Source	在阿根廷首都布宜诺斯艾利斯参加联合国全球气候大会的代表们继续进行工作。
Gloss	*in argentine capital beunos aires participate united nations global climate conference delegates continue to work.*
Ref 0	delegates attending the un conference on world climate continue their work in the argentine capital of buenos aires.
Ref 1	the delegates to the un global climate conference held in Buenos aires, capital city of argentina, go on with their work.
Ref 2	the delegates continue their works at the united nations global climate talks in buenos aires, capital of argentina
Ref 3	in buenos aires, the capital of argentina, the representatives attending un global climate meeting continued their work.
Full-Hiero: Baseline	in the argentine capital of buenos aires to attend the un conference on global climate of representatives continue to work.
Sh-1 Hiero: R-glue & X-glue	the representatives were in the argentine capital of beunos aires to attend the un conference on global climate continues to work.

Table 1: An example for the level of reordering in Chinese-English translation

Grammar	**Glue:** X^N	**Glue:** X	**Time**
Full Hiero	**25.96**		0.71
Shallow-1	23.54	24.04	0.24
+ R-Glue	23.41	24.15	0.25
+ X-Glue	23.75	**24.74**	0.72
Shallow-2	24.54	25.12	0.55
+ R-Glue	24.75	**25.60**	0.57
+ X-Glue	24.33	25.43	0.69
Shallow-3	24.88	25.89	0.62
+ R-Glue	24.77	**26.24**	0.63
+ X-Glue	24.75	25.83	0.69

Table 2: Results for Chinese-English. The decoding time is in secs/word on the Test set for column 'Glue: X'. Bold font indicate best BLEU for each shallow-order.

Grammar	**Glue:** X	**Time**
Full Hiero	**37.54**	0.67
Shallow-1	36.90	0.40
+ R-Glue	36.98	0.43
+ X-Glue	**37.21**	0.57
Shallow-2	36.97	0.57
+ R-Glue	36.80	0.58
+ X-Glue	**37.36**	0.61
Shallow-3	36.88	0.61
+ R-Glue	37.18	0.63
+ X-Glue	**37.31**	0.64

Table 3: Results for Arabic-English. The decoding time is in secs/word on the Test set.

shallow-3 free glue variants achieve BLEU scores comparable to full-Hiero and at the same time being $12-20\%$ faster.

R-glue (r_g) appears to contribute more than the X-glue (x_g) as can be seen in shallow-2 and shallow-3 cases. Interestingly, x_g is more helpful for the shallow-1 case specifically when the glue rules are restricted. As the glue rules are restricted, the X-glue rules concatenates other lower-order rules before being folded into the glue rules. Both r_g and x_g improve the BLEU scores by 0.58 over the plain shallow case for shallow orders 1 and 2 and performs comparably for shallow-3 case. We have also conducted experiments for Arabic-English (Table 3) and we notice that X-glue is more effective and that R-glue is helpful for higher shallow orders.

5.1 Effect of our novel LM integration

Here we analyze the effect of our novel LM integration approach in terms of BLEU score and search errors comparing it to the naive method used in typical Hiero systems. In shallow setting, our method improved the BLEU scores by 0.4 for both Ar-En and Cn-En. In order to quantify the change in the search errors, we compare the model scores of the (corresponding) candidates in the N-best lists obtained by the two methods and compute the $\%$ of high scoring candidates in each. Our approach was clearly superior with 94.6% and 77.3% of candidates having better scores respectively for Cn-En and Ar-En. In full decoding setting the margin of improvements were reduced slightly- BLEU improved by 0.3 and about $57-69\%$ of target candidates had better model scores for the two language pairs.

References

Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. Bayesian synchronous grammar induction. In *Proceedings of Neural Information Processing Systems*.

Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. 2009. A gibbs sampler for phrasal synchronous grammar induction. In *Proceedings of Association of Computational Linguistics*, pages 782–790.

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

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33.

Adrià de Gispert, Gonzalo Iglesias, Graeme Blackwood, Eduardo R. Banga, and William Byrne. 2010. Hierarchical phrase-based translation with weighted finite-state transducers and shallow-n grammars. *Computational Linguistics*, 36.

Zhongjun He, Yao Meng, and Hao Yu. 2009. Discarding monotone composed rule for hierarchical phrase-based statistical machine translation. In *Proceedings of the 3rd International Universal Communication Symposium*, pages 25–29.

Gonzalo Iglesias, Adrià de Gispert, Eduardo R. Banga, and William Byrne. 2009. Rule filtering by pattern for efficient hierarchical translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 380–388.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of Association of Computational Linguistics*, pages 311–318.

Markus Saers, Joakim Nivre, and Dekai Wu. 2009. Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 29–32. Association for Computational Linguistics.

Baskaran Sankaran, Gholamreza Haffari, and Anoop Sarkar. 2011. Bayesian extraction of minimal scfg rules for hierarchical phrase-based translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 533–541.

Baskaran Sankaran, Majid Razmara, and Anoop Sarkar. 2012. Kriya – an end-to-end hierarchical phrase-based mt system. *The Prague Bulletin of Mathematical Linguistics*, 97(97):83–98, April.

Andreas Zollmann, Ashish Venugopal, Franz Och, and Jay Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical mt. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 1145–1152.