A Systematic Comparison of Phrase-Based, Hierarchical and Syntax-Augmented Statistical MT

Andreas Zollmann* and **Ashish Venugopal*** and **Franz Och** and **Jay Ponte**
Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94303, USA
{zollmann,ashishv}@cs.cmu.edu {och,ponte}@google.com

Abstract

Probabilistic synchronous context-free grammar (PSCFG) translation models define weighted transduction rules that represent translation and reordering operations via nonterminal symbols. In this work, we investigate the source of the improvements in translation quality reported when using two PSCFG translation models (hierarchical and syntax-augmented), when extending a state-of-the-art phrase-based baseline that serves as the lexical support for both PSCFG models. We isolate the impact on translation quality for several important design decisions in each model. We perform this comparison on three NIST language translation tasks; Chinese-to-English, Arabic-to-English and Urdu-to-English, each representing unique challenges.

1 Introduction

Probabilistic synchronous context-free grammar (PSCFG) models define weighted transduction rules that are automatically learned from parallel training data. As in monolingual parsing, such rules make use of nonterminal categories to generalize beyond the lexical level. In the example below, the French (source language) words "ne" and "pas" are translated into the English (target language) word "not", performing reordering in the context of a nonterminal of type "VB" (verb).

$$\text{VP} \rightarrow \text{ne VB pas}, \text{do not VB} : w_1$$

$$\text{VB} \rightarrow \text{veux}, \text{want} : w_2 \ .$$

As with probabilistic context-free grammars, each rule has a left-hand-side nonterminal (VP and VB in the two rules above), which constrains the rule's usage in further composition, and is assigned a weight w, estimating the quality of the rule based on some underlying statistical model. Translation with a PSCFG is thus a process of composing such rules to parse the source language while synchronously generating target language output. PSCFG approaches such as Chiang (2005) and Zollmann and Venugopal (2006) typically begin with a phrase-based model as the foundation for the PSCFG rules described above. Starting with bilingual phrase pairs extracted from automatically aligned parallel text (Och and Ney, 2004; Koehn et al., 2003), these PSCFG approaches augment each contiguous (in source and target words) phrase pair with a left-hand-side symbol (like the VP in the example above), and perform a generalization procedure to form rules that include nonterminal symbols. We can thus view PSCFG methods as an attempt to generalize beyond the purely lexical knowledge represented in phrase based models, allowing reordering decisions to be explicitly encoded in each rule. It is important to note that while phrase-based models cannot explicitly represent context sensitive reordering effects like those in the example above, in practice, phrase based models often have the potential to generate the same target translation output by translating source phrases out of order, and allowing empty translations for some source words. Apart from one or more language models scoring these reordering alternatives, state-of-the-art phrase-based systems are also equipped with a lexicalized distortion model accounting for reordering behavior more directly. While previous work demonstrates impres-

*Work done during internships at Google Inc.

sive improvements of PSCFG over phrase-based approaches for large Chinese-to-English data scenarios (Chiang, 2005; Chiang, 2007; Marcu et al., 2006; DeNeefe et al., 2007), these phrase-based baseline systems were constrained to distortion limits of four (Chiang, 2005) and seven (Chiang, 2007; Marcu et al., 2006; DeNeefe et al., 2007), respectively, while the PSCFG systems were able to operate within an implicit reordering window of 10 and higher.

In this work, we evaluate the impact of the extensions suggested by the PSCFG methods above, looking to answer the following questions. Do the relative improvements of PSCFG methods persist when the phrase- based approach is allowed comparable long-distance reordering, and when the n-gram language model is strong enough to effectively select from these reordered alternatives? Do these improvements persist across language pairs that exhibit significantly different reodering effects and how does resource availablility effect relative performance? In order to answer these questions we extend our PSCFG decoder to efficiently handle the high order LMs typically applied in state-of-the-art phrase based translation systems. We evaluate the phrase-based system for a range of reordering limits, up to those matching the PSCFG approaches, isolating the impact of the nonterminal based approach to reordering. Results are presented in multiple language pairs and data size scenarios, highlighting the limited impact of the PSCFG model in certain conditions.

2 Summary of approaches

Given a source language sentence $\mathbf{f}$, statistical machine translation defines the translation task as selecting the most likely target translation $\mathbf{e}$ under a model $P(\mathbf{e}|\mathbf{f})$, i.e.:

$$\hat{\mathbf{e}}(\mathbf{f}) = \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f}) = \arg\max_{\mathbf{e}} \sum_{i=1}^{m} h_i(\mathbf{e}, \mathbf{f})\lambda_i$$

where the $\arg\max$ operation denotes a search through a structured space of translation ouputs in the target language, $h_i(\mathbf{e}, \mathbf{f})$ are bilingual features of $\mathbf{e}$ and $\mathbf{f}$ and monolingual features of $\mathbf{e}$, and weights λ_i are trained discriminitively to maximize translation quality (based on automatic metrics) on held out data (Och, 2003).

Both phrase-based and PSCFG approaches make independence assumptions to structure this search space and thus most features $h_i(\mathbf{e}, \mathbf{f})$ are designed to be local to each phrase pair or rule. A notable exception is the n-gram language model (LM), which evaluates the likelihood of the sequential target words output. Phrase-based systems also typically allow source segments to be translated out of order, and include distortion models to evaluate such operations. These features suggest the efficient dynamic programming algorithms for phrase-based systems described in Koehn et al. (2004).

We now discuss the translation models compared in this work.

2.1 Phrase Based MT

Phrase-based methods identify contiguous bilingual phrase pairs based on automatically generated word alignments (Och et al., 1999). Phrase pairs are extracted up to a fixed maximum length, since very long phrases rarely have a tangible impact during translation (Koehn et al., 2003). During decoding, extracted phrase pairs are reordered to generate fluent target output. Reordered translation output is evaluated under a distortion model and corroborated by one or more n-gram language models. These models do not have an explicit representation of how to reorder phrases. To avoid search space explosion, most systems place a limit on the distance that source segments can be moved within the source sentence. This limit, along with the phrase length limit (where local reorderings are implicit in the phrase), determine the scope of reordering represented in a phrase-based system. All experiments in this work limit phrase pairs to have source and target length of at most 12, and either source length or target length of at most 6 (higher limits did not result in additional improvements). In our experiments phrases are extracted by the method described in Och and Ney (2004) and reordering during decoding with the lexicalized distortion model from Zens and Ney (2006). The reordering limit for the phrase based system (for each language pair) is increased until no additional improvements result.

2.2 Hierarchical MT

Building upon the success of phrase-based methods, Chiang (2005) presents a PSCFG model of translation that uses the bilingual phrase pairs of phrase-based MT as starting point to learn hierarchical rules. For each training sentence pair's set of extracted phrase pairs, the set of induced PSCFG rules can be generated as follows: First, each

phrase pair is assigned a generic X-nonterminal as left-hand-side, making it an *initial rule*. We can now recursively generalize each already obtained rule (initial or including nonterminals)

$$N \rightarrow f_1 \dots f_m / e_1 \dots e_n$$

for which there is an *initial* rule

$$M \rightarrow f_i \dots f_u / e_j \dots e_v$$

where $1 \leq i < u \leq m$ and $1 \leq j < v \leq n$, to obtain a new rule

$$N \rightarrow f_1^{i-1} X_k f_{u+1}^m / e_1^{j-1} X_k e_{v+1}^n$$

where e.g. f_1^{i-1} is short-hand for $f_1 \dots f_{i-1}$, and where k is an index for the nonterminal X that indicates the one-to-one correspondence between the new X tokens on the two sides (it is not in the space of word indices like i, j, u, v, m, n). The recursive form of this generalization operation allows the generation of rules with multiple nonterminal symbols.

Performing translation with PSCFG grammars amounts to straight-forward generalizations of chart parsing algorithms for PCFG grammars. Adaptations to the algorithms in the presence of n-gram LMs are discussed in (Chiang, 2007; Venugopal et al., 2007; Huang and Chiang, 2007).

Extracting hierarchical rules in this fashion can generate a large number of rules and could introduce significant challenges for search. Chiang (2005) places restrictions on the extracted rules which we adhere to as well. We disallow rules with more than two nonterminal pairs, rules with adjacent source-side nonterminals, and limit each rule's source side length (i.e., number of source terminals and nonterminals) to 6. We extract rules from initial phrases of maximal length 12 (exactly matching the phrase based system).[1] Higher length limits or allowing more than two nonterminals per rule do not yield further improvements for systems presented here.

During decoding, we allow application of all rules of the grammar for chart items spanning up to 15 source words (for sentences up to length 20), or 12 source words (for longer sentences), respectively. When that limit is reached, only a special glue rule allowing monotonic concatenation of hypotheses is allowed. (The same holds for the Syntax Augmented system.)

[1] Chiang (2005) uses source length limit 5 and initial phrase length limit 10.

2.3 Syntax Augmented MT

Syntax Augmented MT (SAMT) (Zollmann and Venugopal, 2006) extends Chiang (2005) to include nonterminal symbols from target language phrase structure parse trees. Each target sentence in the training corpus is parsed with a stochastic parser—we use Charniak (2000))—to produce constituent labels for target spans. Phrases (extracted from a particular sentence pair) are assigned left-hand-side nonterminal symbols based on the target side parse tree constituent spans. Phrases whose target side corresponds to a constituent span are assigned that constituent's label as their left-hand-side nonterminal. If the target span of the phrase does not match a constituent in the parse tree, heuristics are used to assign categories that correspond to partial rewriting of the tree. These heuristics first consider concatenation operations, forming categories such as "NP+V", and then resort to CCG (Steedman, 1999) style "slash" categories such as "NP/NN." or "DT\NP". In the spirit of isolating the additional benefit of syntactic categories, the SAMT system used here also generates a purely hierarchical (single generic nonterminal symbol) variant for each syntax-augmented rule. This allows the decoder to choose between translation derivations that use syntactic labels and those that do not. Additional features introduced in SAMT rules are: a relative frequency estimated probability of the rule given its left-hand-side nonterminal, and a binary feature for the the purely hierachial variants.

3 Large N-Gram LMs for PSCFG decoding

Brants et al. (2007) demonstrate the value of large high-order LMs within a phrase-based system. Recent results with PSCFG based methods have typically relied on significantly smaller LMs, as a result of runtime complexity within the decoder. In this work, we started with the publicly available PSCFG decoder described in Venugopal et al. (2007) and extended it to efficiently use distributed higher-order LMs under the Cube-Pruning decoding method from Chiang (2007). These extensions allow us to verify that the benefits of PSCFG models persist in the presence of large, powerful n-gram LMs.

3.1 Asynchronous N-Gram LMs

As described in Brants et al. (2007), using large distributed LMs requires the decoder to perform

asynchronous LM requests. Scoring n-grams under this distributed LM involves queuing a set of n-gram probability requests, then distributing these requests in batches to dedicated LM servers, and waiting for the resulting probabilities, before accessing them to score chart items. In order to reduce the number of such roundtrip requests in the chart parsing decoding algorithm used for PSCFGs, we batch all n-gram requests for each cell.

This single batched request per cell paradigm requires some adaptation of the Cube-Pruning algorithm. Cube-Pruning is an early pruning technique used to limit the generation of low quality chart items during decoding. The algorithm calls for the generation of N-Best chart items at each cell (across all rules spanning that cell). The n-gram LM is used to score each generated item, driving the N-Best search algorithm of Huang and Chiang (2005) toward items that score well from a translation model *and* language model perspective. In order to accomodate batched asynchronous LM requests, we queue n-gram requests for the top N*K chart items *without the n-gram LM* where K=100. We then generate the top N chart items *with the n-gram LM* once these probabilties are available. Chart items attempted to be generated during Cube-Pruning that would require LM probabilities of n-grams *not* in the queued set are discarded. While discarding these items could lead to search errors, in practice they tend to be poorly performing items that do not affect final translation quality.

3.2 PSCFG Minimal-State Recombination

To effectively compare PSCFG approaches to state-of-the-art phrase-based systems, we must be able to use high order n-gram LMs during PSCFG decoding, but as shown in Chiang (2007), the number of chart items generated during decoding grows exponentially in the the order of the n-gram LM. Maintaining full $n-1$ word left and right histories for each chart item (required to correctly select the $\arg\max$ derivation when considering a n-gram LM features) is prohibitive for $n > 3$.

We note however, that the full $n-1$ left and right word histories are unneccesary to safely compare two competing chart items. Rather, given the sparsity of high order n-gram LMs, we only need to consider those histories that can actually be found in the n-gram LM. This allows significantly more chart items to be recombined during decoding, without additional search error. The n-gram LM implementation described in Brants et al. (2007) indicates when a particular n-gram is not found in the model, and returns a shortened n-gram or ("state") that represents this shortened condition. We use this state to identify the left and right chart item histories, thus reducing the number of equivalence classes per cell.

Following Venugopal et al. (2007), we also calculate an estimate for the quality of each chart item's left state based on the words represented within the state (since we cannot know the target words that might precede this item in the final translation). This estimate is only used during Cube-Pruning to limit the number of chart items generated.

The extensions above allows us to experiment with the same order of n-gram LMs used in state-of-the-art phrase based systems. While experiments in this work include up to 5-gram models, we have succesfully run these PSCFG systems with higher order n-gram LM models as well.

4 Experiments

4.1 Chinese-English and Arabic-English

We report experiments on three data configurations. The first configuration (Full) uses all the data (both bilingual and monolingual) data available for the NIST 2008 large track translation task. The parallel training data comprises of 9.1M sentence pairs (223M Arabic words, 236M English words) for Arabic-English and 15.4M sentence pairs (295M Chinese Words, 336M English words) for Chinese-English. This configuration (for both Chinese-English and Arabic-English) includes three 5-gram LMs trained on the target side of the parallel data (549M tokens, 448M 1..5-grams), the LDC Gigaword corpus (3.7B tokens, 2.9B 1..5-grams) and the Web 1T 5-Gram Corpus (1T tokens, 3.8B 1..5-grams). The second configuration (TargetLM) uses a single language model trained only on the target side of the parallel training text to compare approaches with a relatively weaker n-gram LM. The third configuration is a simulation of a low data scenario (10%TM), where only 10% of the bilingual training data is used, with the language model from the TargetLM configuration. Translation quality is automatically evaluated by the IBM-BLEU metric (Papineni et al., 2002) (case-sensitive, using length of the closest reference translation) on the following publicly

Ch.-En. System \ %BLEU	Dev (MT04)	MT02	MT03	MT05	MT06	MT08	TstAvg
FULL							
Phraseb. reo=4	37.5	38.0	38.9	36.5	32.2	26.2	**34.4**
Phraseb. reo=7	40.2	40.3	41.1	38.5	34.6	27.7	**36.5**
Phraseb. reo=12	41.3*	41.0	41.8	39.4	35.2	27.9	**37.0**
Hier.	41.6*	40.9	42.5	40.3	36.5	28.7	**37.8**
SAMT	41.9*	41.0	43.0	40.6	36.5	29.2	**38.1**
TARGET-LM							
Phraseb. reo=4	35.9*	36.0	36.0	33.5	30.2	24.6	**32.1**
Phraseb. reo=7	38.3*	38.3	38.6	35.8	31.8	25.8	**34.1**
Phraseb. reo=12	39.0*	38.7	38.9	36.4	33.1	25.9	**34.6**
Hier.	38.1*	37.8	38.3	36.0	33.5	26.5	**34.4**
SAMT	39.9*	39.8	40.1	36.6	34.0	26.9	**35.5**
TARGET-LM, 10%TM							
Phraseb. reo=12	36.4*	35.8	35.3	33.5	29.9	22.9	**31.5**
Hier.	36.4*	36.5	36.3	33.8	31.5	23.9	**32.4**
SAMT	36.5*	36.1	35.8	33.7	31.2	23.8	**32.1**
Ar.-En. System \ %BLEU	Dev (MT04)	MT02	MT03	MT05	MT06	MT08	TstAvg
FULL							
Phraseb. reo=4	51.7	64.3	54.5	57.8	45.9	44.2	**53.3**
Phraseb. reo=7	51.7*	64.5	54.3	58.2	45.9	44.0	**53.4**
Phraseb. reo=9	51.7	64.3	54.4	58.3	45.9	44.0	**53.4**
Hier.	52.0*	64.4	53.5	57.5	45.5	44.1	**53.0**
SAMT	52.5*	63.9	54.2	57.5	45.5	44.9	**53.2**
TARGET-LM							
Phraseb. reo=4	49.3	61.3	51.4	53.0	42.6	40.2	**49.7**
Phraseb. reo=7	49.6*	61.5	51.9	53.2	42.8	40.1	**49.9**
Phraseb. reo=9	49.6	61.5	52.0	53.4	42.8	40.1	**50.0**
Hier.	49.1*	60.5	51.0	53.5	42.0	40.0	**49.4**
SAMT	48.3*	59.5	50.0	51.9	41.0	39.1	**48.3**
TARGET-LM, 10%TM							
Phraseb. reo=7	47.7*	59.4	50.1	51.5	40.5	37.6	**47.8**
Hier.	46.7*	58.2	48.8	50.6	39.5	37.4	**46.9**
SAMT	45.9*	57.6	48.7	50.7	40.0	37.3	**46.9**

Table 1: Results (% case-sensitive IBM-BLEU) for Ch-En and Ar-En NIST-large. Dev. scores with * indicate that the parameters of the decoder were MER-tuned for this configuration and also used in the corresponding non-marked configurations.

available NIST test corpora: MT02, MT03, MT05, MT06, MT08. We used the NIST MT04 corpus as development set to train the model parameters λ. All of the systems were evaluated based on the argmax decision rule. For the purposes of stable comparison across multiple test sets, we additionally report a TstAvg score which is the average of all test set scores.[2]

Table 1 shows results comparing phrase-based, hierarchical and SAMT systems on the Chinese-English and Arabic-English large-track NIST 2008 tasks. Our primary goal in Table 1 is to evaluate the relative impact of the PSCFG methods above the phrase-based approach, and to verify that these improvements persist with the use of of large n-gram LMs. We also show the impact of larger reordering capability under the phrase-based approach, providing a fair comparison to the PSCFG approaches.

[2]We prefer this over taking the average over the aggregate test data to avoid artificially generous BLEU scores due to length penalty effects resulting from e.g. being too brief in a hard test set but compensating this by over-generating in an easy test set.

Chinese-to-English configurations: We see consistent improvements moving from phrase-based models to PSCFG models. This trend holds in both LM configurations (Full and TargetLM) as well as the 10%TM case, with the exception of the hierarchical system for TargetLM, which performs slightly worse than the maximum-reordering phrase-based system.

We vary the reordering limit "reo" for the phrase-based Full and TargetLM configurations and see that Chinese-to-English translation requires significant reordering to generate fluent translations, as shown by the TstAvg difference between phrase-based reordering limited to 4 words (34.4) and 12 words (37.0). Increasing the reordering limit beyond 12 did not yield further improvement. Relative improvements over the most capable phrase-based model demonstrate that PSCFG models are able to model reordering effects more effectively than our phrase-based approach, even in the presence of strong n-gram LMs (to aid the distortion models) and comparable reordering constraints.

Our results with hierarchical rules are consistent with those reported in Chiang (2007), where the hierarchical system uses a reordering limit of 10 (implicit in the maximum length of the initial phrase pairs used for the construction of the rules, and the decoder's maximum source span length, above which only the glue rule is applied) and is compared to a phrase-based system with a reordering limit of 7.

Arabic-to-English configurations: Neither the hierarchical nor the SAMT system show consistent improvements over the phrase-based baseline, outperforming the baseline on some test sets, but underperforming on others. We believe this is due to the lack of sufficient reordering phenomena between the two languages, as evident by the minimal TstAvg improvement the phrase-based system can achieve when increasing the reordering limit from 4 words (53.3) to 9 words (53.4).

N-Gram LMs: The impact of using additional language models in configuration Full instead of only a target-side LM (configuration TargetLM) is clear; the phrase-based system improves the TstAvg score from 34.6 to 37.0 for Chinese-English and from 50.0 to 53.4 for Arabic-English. Interestingly, the hierarchical system and SAMT benefit from the additional LMs to the same extent, and retain their relative improvement compared to the phrase-based system for Chinese-English.

Expressiveness: In order to evaluate how much of the improvement is due to the relatively weaker expressiveness of the phrase-based model, we tried to regenerate translations produced by the hierarchical system with the phrase-based decoder by limiting the phrases applied during decoding to those matching the desired translation ('forced translation'). By forcing the phrase-based system to follow decoding hypotheses consistent with a specific target output, we can determine whether the phrase-based system could possibly generate this output. We used the Chinese-to-English NIST MT06 test (1664 sentences) set for this experiment. Out of the hierarchical system's translations, 1466 (88%) were generable by the phrase-based system. The relevant part of a sentence for which the hierarchical translation was not phrase-based generable is shown in Figure 1. The reason for the failure to generate the translation is rather unspectacular: While the hierarchical system is able to delete the Chinese word meaning 'already' using the rule spanning [27-28], which it learned by generalizing a training phrase pair in which 'already' was not explicitly represented in the target side, the phrase-based system has to account for this Chinese word either directly or in a phrase combining the previous word (Chinese for 'epidemic') or following word (Chinese for 'outbreak').

Out of the generable forced translations, 1221 (83%) had a higher cost than the phrase-based system's preferred output; in other words, the fact that the phrase-based system does not prefer these forced translations is mainly inherent in the model rather than due to search errors.

These results indicate that a phrase-based system with sufficiently powerful reordering features and LM might be able to narrow the gap to a hierarchical system.

System \ %BLEU	Dev	MT08
Phr.b. reo=4	12.8	**18.1**
Phr.b. reo=7	14.2	**19.9**
Phr.b. reo=10	14.8*	**20.2**
Phr.b. reo=12	15.0	**20.1**
Hier.	16.0*	**22.1**
SAMT	16.1*	**22.6**

Table 2: Translation quality (% case-sensitive IBM-BLEU) for Urdu-English NIST-large. We mark dev. scores with * to indicate that the parameters of the corresponding decoder were MER-tuned for this configuration.

4.2 Urdu-English

Table 2 shows results comparing phrase-based, hierarchical and SAMT system on the Urdu-English large-track NIST 2008 task. Systems were trained on the bilingual data provided by the NIST competition (207K sentence pairs; 2.2M Urdu words / 2.1M English words) and used a n-gram LM estimated from the English side of the parallel data (4M 1..5-grams). We see clear improvements moving from phrase-based to hierarchy, and additional improvements from hierarchy to syntax. As with Chinese-to-English, longer-distance reordering plays an important role when translating from Urdu to English (the phrase-based system is able to improve the test score from 18.1 to 20.2), and PSCFGs seem to be able to take this reordering better into account than the phrasal distance-based and lexical-reordering models.

4.3 Are all rules important?

One might assume that only a few hierarchical rules, expressing reordering phenomena based on common words such as prepositions, are sufficient to obtain the desired gain in translation quality

```
Source:      ... 怀疑 (suspect) 疫情 (epidemic) 已经 (already) 爆发 (outbreak) 的 ('s) 消息 (news) .
Reference:   ... news ... about suspicions of breakouts of the epidemic.
Phasebased: ... the epidemic has already broke the news of doubts.
Hierarchical system:
        ( [26-32: @X -> @X 的 @X . / the @X^2 of the @X^1 .]   the
            ( [31-31: @X -> 消息 / news]   news
            ) of the
            ( [26-29: @X -> 怀疑 @X / suspected @X^1]   suspected
                ( [27-29: @X -> @X 爆发 / outbreak of the @X^1]   outbreak of the
                    ( [27-28: @X -> @X 已经 / @X^1]
                        ( [27-27: @X -> 疫情 / epidemic]   epidemic
                        )
                    )
                )
            ) .
        )
```

Figure 1: Example from NIST MT06 for which the hierarchical system's first best hypothesis was not generable by the phrase-based system. The hierarchical system's decoding parse tree contains the translation in its leaves in infix order (shaded). Each non-leaf node denotes an applied PSCFG rule of the form: [Spanned-source-positions:Left-hand-side->source/target]

Ch.-En. System \ %BLEU	Dev (MT04)	MT02	MT03	MT05	MT06	MT08	TstAvg
Phraseb.	41.3*	41.0	41.8	39.4	35.2	27.9	**37.0**
Hier. default (mincount=3)	41.6*	40.9	42.5	40.3	36.5	28.7	**37.8**
Hier. mincount=4	41.4	41.0	42.5	40.4	36.1	28.4	**37.7**
Hier. mincount=8	41.0	41.0	42.0	40.5	35.7	27.8	**37.4**
Hier. mincount=16	40.7	40.3	41.5	40.0	35.2	27.8	**37.0**
Hier. mincount=32	40.4	40.0	41.5	39.5	34.8	27.5	**36.6**
Hier. mincount=64	39.8	40.0	40.9	39.1	34.6	27.3	**36.4**
Hier. mincount=128	39.4	39.8	40.3	38.7	34.0	26.6	**35.9**
Hier. 1NT	40.1*	39.8	41.1	39.1	35.1	28.1	**36.6**

Urdu-En. System \ %BLEU	Dev	MT08
Phraseb.	15.0*	**20.1**
Hier. default (mincount=2)	16.0*	**22.1**
Hier. mincount=4	15.7	**22.0**
Hier. mincount=8	15.4	**21.5**
Hier. mincount=16	15.1	**21.3**
Hier. mincount=32	14.9	**20.7**
Hier. mincount=64	14.6	**20.1**
Hier. mincount=128	14.4	**19.6**
Hier. 1NT	15.3*	**20.8**

Table 3: Translation quality (% case-sensitive IBM-BLEU) for Chinese-English and Urdu-English NIST-large when restricting the hierarchical rules. We mark dev. scores with * to indicate that the parameters of the corresponding decoder were MER-tuned for this configuration.

over a phrase-based system. Limiting the number of rules used could reduce search errors caused by spurious ambiguity during decoding. Potentially, hierarchical rules based on rare phrases may not be needed, as these phrase pairs can be substituted into the nonterminal spots of more general and more frequently encountered hierarchical rules.

As Table 3 shows, this is not the case. In these experiments for Hier., we retained all non-hierarchical rules (i.e., phrase pairs) but removed hierarchical rules below a threshold 'mincount'. Increasing mincount to 16 (Chinese-English) or 64 (Urdu-English), respectively, already deteriorates performance to the level of the phrase-based system, demonstrating that the highly parameterized reordering model implicit in using more rules does result in benefits. This immediate reduction in translation quality when removing rare rules can be explained by the following effect. Unlike in a phrase-based system, where any phrase can potentially be reordered, rules in the PSCFG must compose to generate sub-translations that can be reordered. Removing rare rules, even those that are highly lexicalized and do not perform any reordering (but still include nonterminal symbols), increases the likelihood that the glue rule is applied simply concatenating span translations without reordering.

Removing hierarchical rules occurring at most twice (Chinese-English) or once (Urdu-English), respectively, did not impact performance, and led to a significant decrease in rule table size and decoding speed.

We also investigate the relative impact of the

rules with two nonterminals, over using rules with a single nonterminal. Using two nonterminals allows more lexically specific reordering patterns at the cost of decoding runtime. Configuration "Hier. 1NT" represents a hierarchical system in which only rules with at most one nonterminal pair are extracted instead of two as in Configuration "Hier. default". The resulting test set score drop is more than one BLEU point for both Chinese-to-English and Urdu-to-English.

5 Conclusion

In this work we investigated the value of PSCFG approaches built upon state-of-the-art phrase-based systems. Our experiments show that PSCFG approaches can yield substantial benefits for language pairs that are sufficiently non-monotonic. Suprisingly, the gap (or non-gap) between phrase-based and PSCFG performance for a given language pair seems to be consistent across small and large data scenarios, and for weak and strong language models alike. In sufficiently non-monotonic languages, the relative improvements of phrase-based systems persist when compared against a state-of-the art phrase-based system that is capable of equally long reordering operations modeled by a lexicalized distortion model and a strong n-gram language model. We hope that this work addresses several of the important questions that the research community has regarding the impact and value of these PSCFG approaches.

Acknowledgments

We thank Richard Zens and the anonymous reviewers for their useful comments and suggestions.

References

Brants, Thorsten, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proc. of EMNLP-CoNLL.*

Charniak, Eugene. 2000. A maximum entropy-inspired parser. In *Proc. of HLT/NAACL.*

Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proc. of ACL.*

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

DeNeefe, Steve, Kevin Knight, Wei Wang, and Daniel Marcu. 2007. What can syntax-based MT learn from phrase-based MT? In *Proc. of EMNLP-CoNLL.*

Huang, Liang and David Chiang. 2005. Better k-best parsing. In *Proc. of IWPT.*

Huang, Liang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proc. of ACL.*

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of HLT/NAACL.*

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2004. Pharaoh: A beam search decoder for phrase-base statistical machine translation models. In *Proc. of AMTA*.

Marcu, Daniel, Wei Wang, Abdessamad Echihabi, and Kevin Knight. 2006. SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. In *Proc. of EMNLP*.

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

Och, Franz Josef, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. of EMNLP*.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL.*

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL.*

Steedman, Mark. 1999. Alternative quantifier scope in CCG. In *Proc. of ACL.*

Venugopal, Ashish, Andreas Zollmann, and Vogel Stephan. 2007. An efficient two-pass approach to synchronous-CFG driven statistical MT. In *Proc. of HLT/NAACL.*

Zens, Richard and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. In *Proc. of the Workshop on Statistical Machine Translation, HLT/NAACL.*

Zollmann, Andreas and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proc. of the Workshop on Statistical Machine Translation, HLT/NAACL.*