New Parameterizations and Features for PSCFG-Based Machine Translation

Andreas Zollmann Stephan Vogel
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{zollmann,vogel}@cs.cmu.edu

Abstract

We propose several improvements to the hierarchical phrase-based MT model of Chiang (2005) and its syntax-based extension by Zollmann and Venugopal (2006). We add a source-span variance model that, for each rule utilized in a probabilistic synchronous context-free grammar (PSCFG) derivation, gives a confidence estimate in the rule based on the number of source words spanned by the rule and its substituted child rules, with the distributions of these source span sizes estimated during training time.

We further propose different methods of combining hierarchical and syntax-based PSCFG models, by merging the grammars as well as by interpolating the translation models.

Finally, we compare syntax-augmented MT, which extracts rules based on target-side syntax, to a corresponding variant based on source-side syntax, and experiment with a model extension that jointly takes source and target syntax into account.

1 Introduction

The Probabilistic Synchronous Context Free Grammar (PSCFG) formalism suggests an intuitive approach to model the long-distance and lexically sensitive reordering phenomena that often occur across language pairs considered for statistical machine translation. As in monolingual parsing, nonterminal symbols in translation rules are used to generalize beyond purely lexical operations. *Labels* on these nonterminal symbols are often used to enforce syntactic constraints in the generation of bilingual sentences and imply conditional independence assumptions in the statistical translation model. Several techniques have been recently proposed to automatically identify and estimate parameters for PSCFGs (or related synchronous grammars) from parallel corpora (Galley et al., 2004; Chiang, 2005; Zollmann and Venugopal, 2006; Liu et al., 2006; Marcu et al., 2006).

In this work, we propose several improvements to the hierarchical phrase-based MT model of Chiang (2005) and its syntax-based extension by Zollmann and Venugopal (2006). We add a source span variance model that, for each rule utilized in a probabilistic synchronous context-free grammar (PSCFG) derivation, gives a confidence estimate in the rule based on the number of source words spanned by the rule and its substituted child rules, with the distributions of these source span sizes estimated during training (i.e., rule extraction) time.

We further propose different methods of combining hierarchical and syntax-based PSCFG models, by merging the grammars as well as by interpolating the translation models.

Finally, we compare syntax-augmented MT, which extracts rules based on target-side syntax, to a corresponding variant based on source-side syntax, and experiment with a model extension based on source *and* target syntax.

We evaluate the different models on the NIST large resource Chinese-to-English translation task.

2 Related work

Chiang et al. (2008) introduce *structural distortion features* into a hierarchical phrase-based model, aimed at modeling nonterminal reordering given source span length, by estimating for each possible source span length ℓ a Bernoulli distribution $p(R|\ell)$ where R takes value one if reordering takes place and zero otherwise. Maximum-likelihood estimation of the distribution amounts to simply counting the relative frequency of nonterminal reorderings over all extracted rule instances that incurred a substitution of span length ℓ. In a more fine-grained approach they add a separate binary feature $\langle R, \ell \rangle$ for each combination of reordering truth value R and span length ℓ (where all $\ell \geq 10$ are merged into a single value), and then tune the feature weights discriminatively on a development set. Our approach differs from Chiang et al. (2008) in that we estimate one source span length distribution for each substitution site of each grammar rule, resulting in unique distributions for each rule, estimated from all instances of the rule in the training data. This enables our model to condition reordering range on the individual rules used in a derivation, and even allows to distinguish between two rules r_1 and r_2 that both reorder arguments with identical mean span lengths ℓ, but where the span lengths encountered in extracted instances of r_1 are all close to ℓ, whereas span length instances for r_2 vary widely.

Chen and Eisele (2010) propose a hypbrid approach between hierarchical phrase based MT and a rule based MT system, reporting improvement over each individual model on an English-to-German translation task. Essentially, the rule based system is converted to a single-nonterminal PSCFG, and hence can be combined with the hierarchical model, another single-nonterminal PSCFG, by taking the union of the rule sets and augmenting the feature vectors, adding zero-values for rules that only exist in one of the two grammars. We face the challenge of combining the single-nonterminal hierarchical grammar with a multi-nonterminal syntax-augmented grammar. Thus one hierarchical rule typically corresponds to many syntax-augmented rules. The SAMT system used by Zollmann et al. (2008) adds hierarchical rules separately to the syntax-augmented grammar, resulting in a backbone grammar of well-estimated hierarchical rules supporting the sparser syntactic rules. They allow the model preference between hierarchical and syntax rules to be learned from development data by adding an indicator feature to all rules, which is one for hierarchical rules and zero for syntax rules. However, no empirical comparison is given between the purely syntax-augmented and the hybrid grammar. We aim to fill this gap by experimenting with both models, and further refine the hybrid approach by adding interpolated probability models to the syntax rules.

Chiang (2010) augments a hierarchical phrase-based MT model with binary syntax features representing the source and target syntactic constituents of a given rule's instantiations during training, thus taking source and target syntax into account while avoiding the data-sparseness and decoding-complexity problems of multi-nonterminal PSCFG models. In our approach, the source- and target-side syntax directly determines the grammar, resulting in a nonterminal set derived from the labels underlying the source- and target-language treebanks.

3 PSCFG-based translation

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 outputs 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 typically trained discriminatively to maximize translation quality (based on automatic metrics) on held out data, e.g., using minimum-error-rate training (MERT) (Och, 2003).

In PSCFG-based systems, the search space is structured by automatically extracted rules that model both translation and re-ordering operations.

Most large scale systems approximate the search above by simply searching for the most likely derivation of rules, rather than searching for the most likely translated output. There are efficient algorithms to perform this search (Kasami, 1965; Chappelier and Rajman, 1998) that have been extended to efficiently integrate n-gram language model features (Chiang, 2007; Venugopal et al., 2007; Huang and Chiang, 2007; Zollmann et al., 2008; Petrov et al., 2008).

In this work we experiment with PSCFGs that have been automatically learned from word-aligned parallel corpora. PSCFGs are defined by a source terminal set (source vocabulary) $\mathcal{T}_S$, a target terminal set (target vocabulary) $\mathcal{T}_T$, a shared nonterminal set $\mathcal{N}$ and rules of the form: $X \rightarrow \langle \gamma, \alpha, w \rangle$ where

- $X \in \mathcal{N}$ is a labeled nonterminal referred to as the left-hand-side of the rule.
- $\gamma \in (\mathcal{N} \cup \mathcal{T}_S)*$ is the source side of the rule.
- $\alpha \in (\mathcal{N} \cup \mathcal{T}_T)*$ is the target side of the rule.
- $w \in [0, \infty)$ is a non-negative real-valued weight assigned to the rule; in our model, w is the exponential function of the inner product of features h and weights λ.

3.1 Hierarchical phrase-based 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 pairs.

Chiang (2005) uses features analogous to the ones used in phrase-based translation: a language model neg-log probability, a 'rule given source-side' neg-log-probability, a 'rule given target-side' neg-log-probability, source- and target conditioned 'lexical' neg-log-probabilities based on word-to-word co-occurrences (Koehn et al., 2003), as well as rule, target word, and glue operation counters. We follow Venugopal and Zollmann (2009) to further add a rareness penalty,

$$1/\operatorname{count}(r)$$

where $\operatorname{count}(r)$ is the occurrence count of rule r in the training corpus, allowing the system to learn penalization of low-frequency rules, as well as three indicator features firing if the rule has one, two unswapped, and two swapped nonterminal pairs, respectively.[1]

3.2 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 to produce constituent labels for target spans. Phrase pairs (extracted from a particular sentence pair) are assigned left-hand-side nonterminal symbols based on the target side parse tree constituent spans.

Phrase pairs whose target side corresponds to a constituent span are assigned that constituent's label as their left-hand-side nonterminal. If the target side of the phrase pair is not spanned by a single constituent in the corresponding parse tree, we use the labels of subsuming, subsumed, and neighboring parse tree constituents to assign

[1] Penalization or reward of purely-lexical rules can be indirectly learned by trading off these features with the rule counter feature.

an extended label of the form $C_1 + C_2$, C_1/C_2, or $C_2\backslash C_1$ (the latter two being motivated from the operations in combinatory categorial grammar (CCG) (Steedman, 2000)), indicating that the phrase pair's target side spans two adjacent syntactic categories (e.g., *she went: NP+VB*), a partial syntactic category C_1 missing a C_2 at the right (e.g., *the great: NP/NN*), or a partial C_1 missing a C_2 at the left (e.g., *great wall: DT\NP*), respectively. The label assignment is attempted in the order just described, i.e., assembling labels based on '+' concatenation of two subsumed constituents is preferred, as smaller constituents tend to be more accurately labeled. If no label is assignable by either of these three methods, a default label 'FAIL' is assigned.

In addition to the features used in hierarchical phrase-based MT, SAMT introduces a relative-frequency estimated probability of the rule given its left-hand-side nonterminal.

4 Modeling Source Span Length of PSCFG Rule Substitution Sites

Extracting a rule with k right-hand-side nonterminal pairs, i.e., substitution sites, (from now on called *order-k rule*) by the method described in Section 3 involves $k + 1$ phrase pairs: one phrase pair used as initial rule and k phrase pairs that are sub phrase pairs of the first and replaced by nonterminal pairs. Conversely, during translation, applying this rule amounts to combining k hypotheses from k different chart cells, each represented by a source span and a nonterminal, to form a new hypothesis and file it into a chart cell. Intuitively, we want the source span lengths of these $k + 1$ chart cells to be close to the source side lengths of the $k+1$ phrase pairs from the training corpus that were involved in extracting the rule. Of course, each rule generally was extracted from multiple training corpus locations, with different involved phrase pairs of different lengths. We therefore model $k + 1$ source span length distributions for each order-k rule in the grammar.

Ignoring the discreteness of source span length for the sake of easier estimation, we assume the distribution to be log-normal. This is motivated by the fact that source span length is positive and that we expect its deviation between instances of the same rule to be greater for long phrase pairs than for short ones.

We can now add $\hat{k} + 1$ features to the translation framework, where $\hat{k}$ is the maximum number of PSCFG rule nonterminal pairs, in our case two. Each feature is computed during translation time. Ideally, it should represent the probability of the hypothesized rule given the respective chart cell span length. However, as each competing rule underlies a different distribution, this would require a Bayesian setting, in which priors over distributions are specified. In this preliminary work we take a simpler approach: Based on the rule's span distribution, we compute the probability that a span length no likelier than the one encountered was generated from the distribution. This probability thus yields a confidence estimate for the rule. More formally, let μ be the mean and σ the standard deviation of the logarithm of the span length random variable X concerned, and let x be the span length encountered during decoding. Then the computed confidence estimate is given by

$$P\left(|\ln(X) - \mu| \geq |\ln(x) - \mu|\right) = 2 * Z\left(-(|\ln(x) - \mu|)/\sigma\right)$$

where Z is the cumulative density function of the normal distribution with mean zero and variance one.

The confidence estimate is one if the encountered span length is equal to the mean of the distribution, and decreases as the encountered span length deviates further from the mean. The severity of that decline is determined by the distribution variance: the higher the variance, the less a deviation from the mean is penalized.

Mean and variance of log source span length are sufficient statistics of the log-normal distribution. As we extract rules in a distributed fashion, we use a straightforward parallelization of the online algorithm of Welford (1962) and its improvement by West (1979) to compute the sample variance over all instances of a rule.

5 Merging a Hierarchical and a Syntax-Based Model

While syntax-based grammars allow for more refined statistical models and guide the search by constraining substitution possibilitites in a grammar derivation, grammar sizes tend to be much greater than for hierarchical grammars. Therefore the average occurrence count of a syntax rule is much lower than that of a hierarchical rule, and thus estimated probabilitites are less reliable.

We propose to augment the syntax-based "rule given source side" and "rule given target side" distributions by hierarchical counterparts obtained by marginalizing over the left-hand-side and right-hand-side rule nonterminals. For example, the hierarchical equivalent of the "rule given source side" probability is obtained by summing occurrence counts over all rules that have the same source and target terminals and substitution positions but possibly differ in the left- and/or right-hand side nonterminal labels, divided by the sum of occurrence counts of all rules that have the same source side terminals and source side substitution positions. Similarly, an alternative rareness penalty based on the combined frequency of all rules with the same terminals and substitution positions is obtained.

Using these syntax and hierarchical features side by side amounts to interpolation of the respective probability models in log-space, with minimum-error-rate training (MERT) determining the optimal interpolation coefficient. We also add respective models interpolated with coefficient .5 in probability-space as additional features to the system.

We further experiment with adding hierarchical rules separately to the syntax-augmented grammar, as proposed in Zollmann et al. (2008), with the respective syntax-specific features set to zero. A 'hierarchical-indicator' feature is added to all rules, which is one for hierarchical rules and zero for syntax rules, allowing the joint model to trade of hierarchical against syntactic rules. During translation, the hierarchical and syntax worlds are bridged by glue rules, which allow monotonic concatenation of hierarchical and syntactic partial sentence hypotheses. We separate the glue feature used in hierarchical and syntax-augmented translation into a glue feature that only fires when a hierarchical rule is glued, and a distinct glue feature firing when gluing a syntax-augmented rule.

6 Extension of SAMT to a bilingually parsed corpus

Syntax-based MT models have been proposed both based on target-side syntactic annotations (Galley et al., 2004; Zollmann and Venugopal, 2006) as well source-side annotations (Liu et al., 2006). Syntactic annotations for both source *and* target language are available for popular language pairs such as Chinese-English. In this case, our grammar extraction procedure can be easily extended to impose both source and target constraints on the eligible substitutions simultaneously.

Let N_f be the nonterminal label that would be assigned to a given initial rule when utilizing the source-side parse tree, and N_e the assigned label according to the target-side parse. Then our bilingual model assigns '$N_f + N_e$' to the initial rule. The extraction of complex rules proceeds as before. The number of nonterminals in this model, based on a source-model label set of size s and a target label set of size t, is thus given by st.

7 Experiments

We evaluate our approaches by comparing translation quality according to the IBM-BLEU (Papineni et al., 2002) metric on the NIST Chinese-to-English translation task using MT04 as development set to train the model parameters λ, and MT05, MT06 and MT08 as test sets.

We perform PSCFG rule extraction and decoding using the open-source "SAMT" system (Venugopal and Zollmann, 2009), using the provided implementations for the hierarchical and syntax-augmented grammars. For all systems, we use the bottom-up chart parsing decoder implemented in the SAMT toolkit with a reordering limit of 15 source words, and correspondingly extract rules from initial phrase pairs of maximum source length 15. All rules have at most two nonterminal symbols, which must be non-consecutive on the source side, and rules must contain at least

one source-side terminal symbol.

For parameter tuning, we use the L_0-regularized minimum-error-rate training tool provided by the SAMT toolkit.

The parallel training data comprises of 9.6M sentence pairs (206M Chinese Words, 228M English words). The source and target language parses for the syntax-augmented grammar were generated by the Stanford parser (Klein and Manning, 2003).

The results are given in Table 1. The source span models (indicated by +span) achieve small test set improvements of 0.15 BLEU points on average for the hierarchical and 0.26 BLEU points for the syntax-augmented system, but these are not statistically significant.

Augmenting a syntax-augmented grammar with hierarchical features ("Syntax+hiermodels") results in average test set improvements of 0.5 BLEU points. These improvements are not statistically significant either, but persist across all three test sets. This demonstrates the benefit of more reliable feature estimation. Further augmenting the hierarchical rules to the grammar ("Syntax+hiermodels+hierrules") does not yield additional improvements.

The use of bilingual syntactic parses ('Syntax/src&tgt') turns out detrimental to translation quality. We assume this is due to the huge number of nonterminals in these grammars and the great amount of badly-estimated low-occurrence-count rules. Perhaps merging this grammar with a regular syntax-augmented grammar could yield better results.

We also experimented with a source-parse based model ('Syntax/src'). While not being able to match translation quality of its target-based counterpart, the model still outperforms the hierarchical system on all test sets.

8 Conclusion

We proposed several improvements to the hierarchical phrase-based MT model of Chiang (2005) and its syntax-based extension by Zollmann and Venugopal (2006). We added a source span length model that, for each rule utilized in a probabilistic synchronous context-free grammar (PSCFG) derivation, gives a confidence estimate in the rule based on the number of source words spanned by the rule and its substituted child rules, resulting in small improvements for hierarchical phrase-based as well as syntax-augmented MT.

We further demonstrated the utility of combining hierarchical and syntax-based PSCFG models and grammars.

Finally, we compared syntax-augmented MT, which extracts rules based on target-side syntax, to a corresponding variant based on source-side syntax, showing that target syntax is more benefitial, and unsuccessfully experimented with a model extension that jointly takes source and target syntax into account.

Hierarchical phrase-based MT suffers from spurious ambiguity: A single translation for a given source sentence can usually be accomplished by many different PSCFG derivations. This problem is exacerbated by syntax-augmented MT with its thousands of nonterminals, and made even worse by its joint source-and-target extension. Future research should apply the work of Blunsom et al. (2008) and Blunsom and Osborne (2008), who marginalize over derivations to find the most probable translation rather than the most probable derivation, to these multi-nonterminal grammars.

All source code underlying this work is available under the GNU Lesser General Public License as part of the 'SAMT' system at:
`www.cs.cmu.edu/~zollmann/samt`

Acknowledgements

This work is in part supported by NSF under the Cluster Exploratory program (grant NSF 0844507), and in part by the US DARPA GALE program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF or DARPA.

References

Blunsom, Phil and Miles Osborne. 2008. Probabilistic inference for machine translation. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 215–223, Morristown, NJ, USA. Association for Computational Linguistics.

	Dev (MT04)	MT05	MT06	MT08	**TestAvg**	Time
Hierarchical	38.63	36.51	33.26	25.77	**31.85**	14.3
Hier+span	39.03	36.44	33.29	26.26	**32.00**	16.7
Syntax	39.17	37.17	33.87	26.81	**32.62**	59
Syntax+hiermodels	39.61	37.74	34.30	27.30	**33.11**	68.4
Syntax+hiermodels+hierrules	39.69	37.56	34.66	26.93	**33.05**	34.6
Syntax+span+hiermodels+hierrules	39.81	38.02	34.50	27.41	**33.31**	39.6
Syntax/src+span+hiermodels+hierrules	39.62	37.25	33.99	26.44	**32.56**	20.1
Syntax/src&tgt+span+hiermodels+hierrules	39.15	36.92	33.70	26.24	**32.29**	17.5

Table 1: Translation quality in % case-insensitive IBM-BLEU (i.e., brevity penalty based on closest reference length) for different systems on Chinese-English NIST-large translation tasks. 'TestAvg' shows the average score over the three test sets. 'Time' is the average decoding time per sentence in seconds on one CPU.

Blunsom, Phil, Trevor Cohn, and Miles Osborne. 2008. A discriminative latent variable model for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2).

Chappelier, J.C. and M. Rajman. 1998. A generalized CYK algorithm for parsing stochastic CFG. In *Proceedings of Tabulation in Parsing and Deduction (TAPD)*, pages 133–137, Paris.

Chen, Yu and Andreas Eisele. 2010. Hierarchical hybrid translation between english and german. In Hansen, Viggo and Francois Yvon, editors, *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, pages 90–97. EAMT, EAMT, 5.

Chiang, David, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 224–233, Honolulu, Hawaii, October. Association for Computational Linguistics.

Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Chiang, David. 2010. Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1443–1452, Uppsala, Sweden, July. Association for Computational Linguistics.

Galley, Michael, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics Conference (HLT/NAACL)*.

Huang, Liang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kasami, T. 1965. An efficient recognition and syntax-analysis algorithm for context-free languages. Technical report, Air Force Cambridge Research Lab.

Klein, Dan and Christoper Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Koehn, Philipp, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics Conference (HLT/NAACL)*.

Liu, Yang, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*.

Marcu, Daniel, Wei Wang, Abdessamad Echihabi, and Kevin Knight. 2006. SPMT: Statistical machine translation with syntactified target language phrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sydney, Australia.

Och, Franz J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Papineni, Kishore, 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 the Association for Computational Linguistics (ACL)*.

Petrov, Slav, Aria Haghighi, and Dan Klein. 2008. Coarse-to-fine syntactic machine translation using language projections. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Steedman, Mark. 2000. *The Syntactic Process*. MIT Press.

Venugopal, Ashish and Andreas Zollmann. 2009. Grammar based statistical MT on Hadoop: An end-to-end toolkit for large scale PSCFG based MT. *The Prague Bulletin of Mathematical Linguistics*, 91:67–78.

Venugopal, Ashish, Andreas Zollmann, and Stephan Vogel. 2007. An efficient two-pass approach to synchronous-CFG driven statistical MT. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics Conference (HLT/NAACL)*.

Welford, B. P. 1962. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420.

West, D. H. D. 1979. Updating mean and variance estimates: an improved method. *Commun. ACM*, 22(9):532–535.

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

Zollmann, Andreas, Ashish Venugopal, Franz J. Och, and Jay Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In *Proceedings of the Conference on Computational Linguistics (COLING)*.