Forced Derivations for Hierarchical Machine Translation

Stephan Peitz Arne Mauser Joern Wuebker Hermann Ney
Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
<surname>@cs.rwth-aachen.de

ABSTRACT
We present an efficient framework to estimate the rule probabilities for a hierarchical phrase-based statistical machine translation system from parallel data. In previous work, this was done with bilingual parsing. We use a more efficient approach splitting the bilingual parsing into two stages, which allows us to train a hierarchical translation model on larger tasks. Furthermore, we apply leave-one-out to counteract over-fitting and use the expected count from the inside-outside algorithm to prune the rule set. On the WMT12 Europarl German→English and French→English tasks, we improve translation quality by up to 1.0 BLEU and 0.9 TER while simultaneously reducing the rule set to 5% of the original size.

KEYWORDS: statistical machine translation, hierarchical decoding, translation model training, forced derivation.

Proceedings of COLING 2012: Posters, pages 933–942,
COLING 2012, Mumbai, December 2012.

1 Introduction

In hierarchical machine translation, discontinuous phrases with "gaps" are allowed and the model is formalized as a synchronous context-free grammar (SCFG). This grammar consists of bilingual rules, which are based on bilingual standard phrases and discontinuous phrases. Each bilingual rule rewrites a generic non-terminal X into a pair of strings $\tilde{f}$ and $\tilde{e}$ with both terminals and non-terminals in both languages

$$X \rightarrow \langle \tilde{f}, \tilde{e} \rangle. \tag{1}$$

In the following, we denote $\tilde{f}$ as *source side* and $\tilde{e}$ as *target side* of a bilingual rule. Obtaining these rules is based on a heuristic extraction from automatically word-aligned bilingual training data. Just like in the phrase-based approach, all bilingual rules of a sentence pair are extracted given an alignment. The standard phrases are stored as *lexical rules* in the rule set. In addition, whenever a phrase contains a sub-phrase, this sub-phrase is replaced by a generic non-terminal X. With these hierarchical phrases we can define the *hierarchical rules* in the SCFG. However, this extraction method causes two problems. First, this approach does not consider, whether a rule is extracted from a likely alignment or not. The rule probabilities which are in general defined as relative frequencies are computed based on the joint counts $C(X \rightarrow \langle \tilde{f}, \tilde{e} \rangle)$ of a bilingual rule $X \rightarrow \langle \tilde{f}, \tilde{e} \rangle$

$$p_H(\tilde{f}|\tilde{e}) = \frac{C(X \rightarrow \langle \tilde{f}, \tilde{e} \rangle)}{\sum_{\tilde{f}'} C(X \rightarrow \langle \tilde{f}', \tilde{e} \rangle)}. \tag{2}$$

Thus, the probabilities depend only on simple counts from a word alignment. Another issue is the large number of extracted rules which is exponential in sentence length (Lopez, 2008). To reduce the size of the hierarchical translation model, threshold pruning, which is based on the counts of the rules, can be applied. However, this is connected to the first mentioned difficulty. Using these counts to prune the rule set may reduce the rule set size, but the translation quality can get worse (Zens et al., 2012). An alternative is a more consistent pruning regarding the translation process.

In this work, we present an approach to directly estimate the rule probabilities by applying an expectation-maximization (EM) inspired algorithm. The rule probabilities are computed in both translation directions, i.e. source-to-target $p_H(\tilde{f}|\tilde{e})$ and target-to-source $p_H(\tilde{e}|\tilde{f})$, and are combined in a weighted log-linear model with other features to find the best translation. Similar to the classical EM algorithm, our algorithm is divided into an expectation step and a maximization step. For the expectation step, we parse the training data to get all possible synchronous derivations between the source and target sentences. The parsing is done with a two-parse algorithm where separately first the source sentence and then the target sentence is parsed. From the resulting parse tree of the target parse, the used rules are extracted. Both parsing steps are done with the CYK+ parsing algorithm. After parsing, we apply the inside-outside algorithm on the generated target parse tree to compute expected counts for each applied rule. As maximization step, we update the rule probabilities using the expected counts.

On the German→English and French→English Europarl task from *NAACL 2012 Workshop on Statistical Machine Translation* (WMT12), we show that our presented approach improves the translation quality by up to 1.0 BLEU and 0.9 TER while the rule set is reduced by 95% of the original size.

The paper is organized as follows. In the following Section, we give a short overview of related work. In Section 3, we describe our forced derivation step in detail. Finally, we discuss the experimental results in Section 4, followed by a conclusion.

2 Related Work

In recent years, several works have investigated the direct training of the translation model to close the gap between the extraction and the translation process.

In (Marcu and Wong, 2002), a joint probability model is presented which estimates phrase translation probabilities from a parallel corpus. For aligning the phrases and estimating the probabilities, the EM algorithm is applied. In (Birch et al., 2006) the joint probability model is constrained by a word alignment to limit the complexity.

The problem of over-fitting due to the EM algorithm is analyzed in (DeNero et al., 2006) and a solution is proposed in (Wuebker et al., 2010) by applying leave-one-out. We will adopt the leave-one-out method in this work and show that its benefits translate to the hierarchical case.

Another approach to learn from decoding on the training data is presented in (Duan et al., 2012). In this work, a training method based on forced derivation trees is described. This structure is used to train apart from a translation model, a distortion model, a source language model and a rule sequence model. As first step, they verified their method on a phrase-based system. However, this method can be adapted for the hierarchical approach.

Besides these publications about phrase training for the phrase-based approach, several works have been presented for hierarchical machine translation during the past years. In most of these papers the idea of bilingual parsing on parallel corpora is described.

In (Blunsom et al., 2008) a discriminative model using derivations as a hidden variable is presented. In training, they perform a synchronous parsing of the source and target sentences using a modified CYK algorithm over two dimensions with a time complexity of $\mathcal{O}(J^3 I^3)$ where J is the source sentence length and I the target sentence length. Further, the inside-outside algorithm is employed. The experiments were carried out on a subset of the French→English Europarl corpus (170K sentences) and show comparable results. Another observation is that their model improves as they increase the number of parsable training sentences. Starting from this observation, we will apply our approach on a larger training corpus and show, that we improve the translation quality on a recent task.

Bilingual parsing on parallel corpora is also described in (Huang and Zhou, 2009), (Čmejrek et al., 2009) and (Čmejrek and Zhou, 2010). They also use the EM algorithm to recompute the translation probabilities. In order to do that, in (Huang and Zhou, 2009) the EM algorithm for SCFG is introduced. In the maximization step, the expected counts from the inside-outside algorithm are used to update the translation probabilities for non-lexical rules only. Experiments on the Chinese→English IWSLT 2006 task (40K sentences without punctuation and case information) result in a significantly better BLEU score. In (Čmejrek et al., 2009) and (Čmejrek and Zhou, 2010), this work is extended and they report improvement on a subset of the German→English Europarl corpus (300K sentences without punctuation and case information).

In (Heger et al., 2010), a standard hierarchical machine translation system is combined with phrases trained as in (Wuebker et al., 2010). Experimental results on Arabic→English IWSLT and English→German WMT task show improvements in translation quality and motivate to

investigate the impact of phrase training for the hierarchical approach.

In (Dyer, 2010) a synchronous parsing algorithm is introduced that is based on two successive monolingual parses. Instead of performing one bilingual parse for a given sentences pair, a two-parse algorithm is applied. This improves the average run-time. The authors reported speed improvement on the same task as in (Blunsom et al., 2008). We apply this approach to reduce the run-time of our forced derivation procedure.

Compared to previous described approaches for training the hierarchical translation model, we are now able to employ forced derivation on larger task. Furthermore, we estimate the inside-outside probabilities on the target chart only and calculate the expected count for all type of rules. We also apply a threshold pruning on the rule set using the estimated expected counts. This leads to a more consistent pruning and a smaller rule set. Another difference is that we perform leave-one-out to counteract over-fitting. Further, in the forced derivation procedure, we include the log-linear combination of all features which are used in the translation process except for the language model.

3 Forced Derivation

In the forced derivation procedure, the two-parse algorithm and the inside-outside algorithm are the expectation step of the EM-inspired algorithm. During the expectation step, the expected counts are calculated. First, we need all possible synchronous derivations given the parallel training data. From the resulting parse trees, all applied rules are extracted and the expected count of each rule is estimated with the inside-outside algorithm. In general, a bilingual parser parses all parallel sentences of the training data based on the full extracted rule set of the training data (Huang and Zhou, 2009). However, to calculate all parses efficiently, we apply the two-parse algorithm instead of full bilingual parsing. The two-parse algorithm performs two monolingual parses, one on the source language sentence f_1^J and one on the target language sentence e_1^I. Each parse is done using the CYK+ algorithm as described in (Chappelier and Rajman, 1998). The main advantage of the CYK+ algorithm is that it does not require the grammar to be in Chomsky Normal Form and we can use hierarchical translation rules directly. of the grammar in the algorithm itself. Similar to the CYK algorithm, the basic data structure is a chart with $\frac{J(J+1)}{2}$ cells where J is in this case the size of the source sentence f_1^J. The source sentence can be generated by the grammar if the start symbol of the grammar S is found in top cell $(J,1)$, i.e. $S \Rightarrow^* f_1^J$. The time complexity is $\mathcal{O}(J^3)$ and $\mathcal{O}(I^3)$ respectively. The resulting charts have a space complexity of $\mathcal{O}(J^2)$ and $\mathcal{O}(I^2)$ for the representation of all derivations. In the hierarchical approach of (Chiang, 2005), the set of non-terminals consists of a start symbol S and a generic non-terminal X. Furthermore, the number of non-terminals n on the right hand side of the rules is limited to two. In addition to the hierarchical and lexical rules, a rule set $\mathscr{R}$ is extended with an initial rule and a glue rule

$$S \rightarrow \langle X, X \rangle, S \rightarrow \langle SX, SX \rangle. \tag{3}$$

Given $\mathscr{R}$, the parallel training corpus is parsed with the two-parse algorithm. First, we parse the source sentence f_1^J with the source sides of the rules of the given rule set $\mathscr{R}$ and a sentence pair (f_1^J, e_1^I). Note, we ensure that each source sentence can always be parsed by employing several heuristics during the extraction process to get all necessary rules. From the chart we extract the target side of the used rules. This is done by simply traversing from the top cell $(J,1)$ through the chart. Then the non-terminals on the left hand side and the non-terminals of the target side are annotated with the source span of the corresponding non-terminals in the chart. The

annotated rules are stored in a new rule set $\mathscr{R}_t$. Moreover, we annotate applied initial rules and glue rules. Hence, the set of non-terminals now consists of several annotated start symbols and generic non-terminals. In our implementation, the left hand side with the annotated target side of the bilingual rule and a pointer to the corresponding original rule including all features is stored. Further, to keep the time and memory usage low, we limit the number of target sides for each rule in the described extraction step.

In the second pass, the target sentence e_1^I is parsed using the annotated target sides of the bilingual rules in the new rule set $\mathscr{R}_t$. The annotation of non-terminals of the rules in the first parse ensures that rules applied in the target parse cover only one span in the source sentence. The target sentence is generated by the new grammar and the forced derivation procedure is successful, if the start symbol S_1^J is found in cell $(I, 1)$. In contrast to the source parse, it is possible that a target parse is not found due to the fact that we prune necessary target sides as describe before. If a target sentence can not be parsed, we discard the sentence pair. The generated parse tree represents all possible synchronous derivations between the source and target sentence and the used rules are extracted from the chart as we save a pointer to the corresponding source side of the rule. Again, the extraction procedure starts from the top cell $(I, 1)$ and traverses through the chart of the target parse.

3.1 Inside-Outside Algorithm

For the estimation of the rule probabilities we employ the inside-outside algorithm to calculate the expected count for each rule used in the forced derivation step. In the maximization step the expected counts are used to update the rule probabilities. As described in (Čmejrek et al., 2009), we calculate the expected count based on the inside and outside probabilities depending on the number of non-terminals in the rule. Due to the fact that we do not perform full bilingual parsing, we apply the inside-outside algorithm on the target parse only. Considering a sentence pair (f_1^J, e_1^I), the expected count $C_{FD}(r_n)$ is computed for a rule r_n applied in target parse. Note, the expected counts for a rule are summed up over all sentence pairs of the training data. For the maximization step, we use the expected counts $C_{FD}(r_n)$ of a rule $r_n = X \rightarrow \langle \tilde{f}, \tilde{e} \rangle$ to update the rule probability $p_{FD}(\tilde{f}|\tilde{e})$

$$p_{FD}(\tilde{f}|\tilde{e}) = \frac{C_{FD}(X \rightarrow \langle \tilde{f}, \tilde{e} \rangle)}{\sum_{\tilde{f}'} C_{FD}(X \rightarrow \langle \tilde{f}', \tilde{e} \rangle)}. \tag{4}$$

This is also done for the target-to-source translation probability $p_{FD}(\tilde{e}|\tilde{f})$.

3.2 Leave-one-out

Another issue of phrase model training in general is over-fitting. Due to the fact that all rules which are extracted from a sentence pair are used in the forced derivation step, longer rules are often preferred. Even though those long rules only match for a few sentences of the training data and do not generalize very well, they tend to be assigned very high translation probabilities. In (Wuebker et al., 2010) a leave-one-out method is described which counteracts the over-fitting. This method modifies the translation probabilities in the forced derivation step for each sentence pair. The occurrences of a given rule in a sentence pair (f_n, e_n) are subtracted from the rule counts obtained from the full training data resulting in the modified translation

probability

$$p_{l1o,n}(\tilde{f}|\tilde{e}) = \frac{C(X \rightarrow \langle \tilde{f}, \tilde{e} \rangle) - C_n(X \rightarrow \langle \tilde{f}, \tilde{e} \rangle)}{\sum_{\tilde{f}'} C(X \rightarrow \langle \tilde{f}', \tilde{e} \rangle) - C_n(X \rightarrow \langle \tilde{f}', \tilde{e} \rangle)} \quad (5)$$

where $C_n(X \rightarrow \langle \tilde{f}, \tilde{e} \rangle)$ is the count for the rule $X \rightarrow \langle \tilde{f}, \tilde{e} \rangle$, that was extracted from this sentence pair. Singleton rules, which are rules occurring only in one sentence, are handled differently. These rules get a low probability depending on the source and target rule lengths. Note, the non-terminals on the right hand side of the rules are treated such as terminals. Without leave-one-out, the longer rule

$$X \rightarrow \langle \text{Und } [\ldots] \text{ Strafen, It } [\ldots] \text{ should} \rangle \quad (6)$$

is applied. Such long rules are used only in few sentence pairs and will hardly generalize well to unseen test data. Using leave-one-out, three shorter, more general rules are used for generating this part of the sentence pair

$$X \rightarrow \langle \text{Und zwar } X, \text{It } X \rangle, X \rightarrow \langle \text{sollen } X, X \text{ should} \rangle, X \rightarrow \langle \text{derartige Strafen, says that this} \rangle. \quad (7)$$

4 Experiments

Our experiments were carried out on the German→English and French→English Europarl task from the *NAACL 2012 Workshop on Statistical Machine Translation*. For both tasks, we selected parallel sentences according to two criteria: Only sentences of maximum 100 tokens are considered and the ratio of the vocabulary size of a sentence and the number of its tokens is minimum 80% i.e. we remove sentences that have too many repeated words. The German text was further preprocessed by splitting German compound words using the frequency-based method described in (Koehn and Knight, 2003). For the experiments, we used the open source translation toolkit Jane (Vilar et al., 2010), which has been developed at RWTH and is freely available for non-commercial use. We extended the hierarchical phrase-based machine translation system based on (Chiang, 2005) with the two-parse algorithm and the inside-outside algorithm as described in Section 3.

4.1 Experimental Setup

Given the training data, we created a word alignment with GIZA++ (Och and Ney, 2003). The resulting alignment was used to extract the initial rule set. As the initial rule set is extracted heuristically, we name it *heuristic rule set* in the following. In contrast, the produced rule set after the forced derivation procedure is called *learned rule set*. First, we built a baseline system which is a standard hierarchical phrase-based SMT system with ten features in a log-linear model: translation and word lexicon probabilities in both translation directions (source-to-target and target-to-source), rule penalty, word penalty, language model score and three binary features for hierarchical rules. Furthermore, we used the heuristic rule set to perform our proposed forced derivation procedure and initialized the weights of each rule in the EM-inspired algorithm with log-linear combination of all features. We used a standard set of non-optimized parameters for the log-linear combination. We applied length-based leave-one-out as described in Section 3 and compared to a setup without leave-one-out. For all experiments, we used a 4-gram language model with modified Kneser-Ney smoothing which was trained with the SRILM toolkit (Stolcke, 2002). Further, we used the cube prune algorithm (Huang and Chiang, 2007) to

perform the search. The scaling factors of the features were optimized for BLEU (Papineni et al., 2001) on the development set with Minimum Error Rate Training (Och, 2003) on 100-best lists. The performance of the different setup was evaluated on the development (newstest2010) and the test set (newstest2011) using the two metrics BLEU and TER (Snover et al., 2006).

4.2 Experimental Results

The results of our different experiments are presented in Table 2. Our approach is abbreviated to *FD* (forced derivation). First, we did different preliminary experiments on the German→English task. We then applied the best methods on the French→English task to verify our proposed approach.

During the forced derivation procedure, around 93% of the parallel sentences of the German→English corpus and around 97% of French→English corpus were parsed with the two-parse algorithm. The non-parsable sentences were skipped. In general, those are longer sentences, which are misaligned usually caused by liberal or wrong translation. For a batch of 2000 sentences, the parsing took on average 2.5 hours (without rule set loading time) on a single machine.

First, we performed our proposed method with and without leave-one-out (Table 1). The length-based leave-one-out (*lbl1o*) method outperforms forced derivation without leave-one-out in terms of BLEU and is also slightly better than the baseline. Further, we pruned the final learned rule set by dropping all rules which got a summed up expected count lower than a given threshold. The results for different threshold values are shown in Table 1. Discarding such rules seems to improve the translation quality and in addition reduces the size of the rule set. We ran several setups using different thresholds and compared them on the development set. Even the full learned rule set does not contain all rules of the initial rule set. The reason for that is the pruning in the forced derivation procedure and the skipped non-parsable sentences. Note, that the pruning settings are weaker then in the translation process. We tested the best setup (*cutoff 0.1*) on the test translation set and achieved an improvement of 0.4 points in BLEU and 0.3 points in TER. The final rule set size is reduced by more than 95%. It seems that the greatest improvement is achieved by this reduction. The results of the experiment using the heuristic rule set filtered to contain the same rules as the pruned learned rule set (*baseline filtered*) are similar to the setup using the translation probabilities learned with the EM-inspired algorithm. This observation shows that using filtered rules performs as least as good as using the full rule set. However, due to the reduced rule set, following experiments were consuming less computation time and memory.

We achieved further improvement applying a log-linear interpolation of the learned rule set with the heuristic one as proposed in (DeNero et al., 2006). The log-linear interpolations $p_{int}(\tilde{f}|\tilde{e})$ are computed as

$$p_{int}(\tilde{f}|\tilde{e}) = \left(p_H(\tilde{f}|\tilde{e})\right)^{1-\omega} \cdot \left(p_{FD}(\tilde{f}|\tilde{e})\right)^{\omega} \quad (8)$$

where ω is the interpolation weight, p_H the heuristic rule set and p_{FD} the learned rule set. Only the intersection of both tables is retained. The interpolation weight was adjusted on the development set and set to $\omega = 0.2$. Our final result shows an improvement of 0.7 BLEU points and 0.8 TER points over the baseline on the test translation set of the German→English task.

For the the French-English task, we applied forced derivation with length-based leave-one-out and a cutoff threshold of 0.1. Similar to the German→English task, we got an improvement of

cutoff threshold	dev BLEU[%]	% of full rule set	
		all type of rules	hierarchical only
0.2	21.0	3.2	3.0
0.15	21.4	3.9	3.6
0.1	**21.4**	**4.9**	**4.7**
0.01	21.2	13.2	15.0
full (length-based l1o)	21.0	92.0	94.3
full (without l1o)	20.3	92.0	94.3
baseline	20.8	100	100

Table 1: Preliminary experiments on the development set of the German→English WMT12 task.

setup	German→English		French→English	
	BLEU[%]	TER[%]	BLEU[%]	TER[%]
baseline	19.1	63.4	24.6	57.2
baseline (filtered)	19.5	63.3	-	-
FD +lbl1o +cutoff 0.1	19.5	63.1	25.0	57.2
fixed interpolation $\omega = 0.2$	**19.8**	**62.6**	**25.6**	**56.3**

Table 2: Final results for the German→English and French→English WMT12 task.

0.4 points in BLEU while the rule set size was reduced by more than 95%. With the log-linear interpolation, we gained further 0.6 BLEU points. In sum, we achieved an improvement of 1.0 points in BLEU and 0.9 points in TER over the baseline on the French-English task.

Conclusion

In this paper, we have introduced an efficient method to perform the direct estimation of rule probabilities for hierarchical machine translation. Based on an EM-inspired algorithm, the expectation is computed with the two-parse algorithm that generates all possible synchronous derivations between a source and target sentence. We applied the inside-outside algorithm to calculate the expected counts and to estimate rules probabilities. To avoid over-fitting, we used length-based leave-one-out. By pruning rules with a low expected count, it is possible to significantly reduce the rule set size.

Compared to previous work, we have also shown improvements on an medium sized task. On the WMT12 Europarl German→English task we improved translation quality by 0.4 BLEU points with the trained rule set and 0.7 BLEU points using the interpolation. Furthermore, the rule set size was reduced by over 95%. In addition, we showed improvements of up to 1.0 BLEU and 0.9 TER on the WMT12 Europarl French→English task.

In future work, a leave-one-out strategy considering non-terminals in a more sophisticated way could further improve forced derivation.

Acknowledgments

This work was partly funded by the European Union under the FP7 project T4ME Net, Contract nº 249119. The research leading to these results has also received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement nº 287658.

References

Birch, A., Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Constraining the Phrase-Based, Joint Probability Statistical Translation Model. In *Human Language Technology Conf. (HLT-NAACL): Proc. Workshop on Statistical Machine Translation*, pages 154–157, New York City, NY.

Blunsom, P., Cohn, T., and Osborne, M. (2008). A discriminative latent variable model for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 200–208, Columbus, Ohio. Association for Computational Linguistics.

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

Chiang, D. (2005). A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 263–270, Ann Arbor, Michigan.

DeNero, J., Gillick, D., Zhang, J., and Klein, D. (2006). Why generative phrase models underperform surface heuristics. In *Workshop on Statistical Machine Translation at HLT-NAACL*.

Duan, N., Li, M., and Zhou, M. (2012). Forced derivation tree based model training to statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 445–454, Jeju Island, Korea. Association for Computational Linguistics.

Dyer, C. (2010). Two monolingual parses are better than one (synchronous parse). In *In Proc. of HLT-NAACL*.

Heger, C., Wuebker, J., Vilar, D., and Ney, H. (2010). A combination of hierarchical systems with forced alignments from phrase-based systems. In *International Workshop on Spoken Language Translation*, pages 291–297, Paris, France.

Huang, L. and Chiang, D. (2007). Forest rescoring: Faster decoding with integrated language models. *Proceedings of ACL 2007*, 45(1):144.

Huang, S. and Zhou, B. (2009). An em algorithm for scfg in formal syntax-based translation. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4813–4816.

Koehn, P. and Knight, K. (2003). Empirical Methods for Compound Splitting. In *Proc. 10th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL)*, pages 347–354, Budapest, Hungary.

Lopez, A. (2008). Tera-scale translation models via pattern matching. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 505–512, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marcu, D. and Wong, W. (2002). A Phrase-Based, Joint Probability Model for Statistical Machine Translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 133–139, Philadelphia, PA.

Och, F. J. (2003). Minimum Error Rate Training for Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Stolcke, A. (2002). SRILM – An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO.

Čmejrek, M. and Zhou, B. (2010). Two methods for extending hierarchical rules from the bilingual chart parsing. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 180–188, Stroudsburg, PA, USA. Association for Computational Linguistics.

Čmejrek, M., Zhou, B., and Xiang, B. (2009). Enriching scfg rules directly from efficient bilingual chart parsing. In *IWSLT'09*, pages 136–143.

Vilar, D., Stein, D., Huck, M., and Ney, H. (2010). Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 262–270, Uppsala, Sweden.

Wuebker, J., Mauser, A., and Ney, H. (2010). Training phrase translation models with leaving-one-out. In *Proceedings of the 48th Annual Meeting of the Assoc. for Computational Linguistics*, pages 475–484, Uppsala, Sweden.

Zens, R., Stanton, D., and Xu, P. (2012). A systematic comparison of phrase table pruning techniques. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 972–983, Jeju Island, Korea.