The RWTH Aachen Machine Translation System for WMT 2012

Matthias Huck, Stephan Peitz, Markus Freitag, Malte Nuhn and 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

This paper describes the statistical machine translation (SMT) systems developed at RWTH Aachen University for the translation task of the *NAACL 2012 Seventh Workshop on Statistical Machine Translation* (WMT 2012). We participated in the evaluation campaign for the French-English and German-English language pairs in both translation directions. Both hierarchical and phrase-based SMT systems are applied. A number of different techniques are evaluated, including an insertion model, different lexical smoothing methods, a discriminative reordering extension for the hierarchical system, reverse translation, and system combination. By application of these methods we achieve considerable improvements over the respective baseline systems.

1 Introduction

For the WMT 2012 shared translation task[1] RWTH utilized state-of-the-art phrase-based and hierarchical translation systems as well as an in-house system combination framework. We give a survey of these systems and the basic methods they implement in Section 2. For both the French-English (Section 3) and the German-English (Section 4) language pair, we investigate several different advanced techniques. We concentrate on specific research directions for each of the translation tasks and present the respective techniques along with the empirical results they yield: For the French→English task (Section 3.1), we apply a standard phrase-based system. For the English→French task (Section 3.2), we augment a hierarchical phrase-based setup with a number of enhancements like an insertion model, different lexical smoothing methods, and a discriminative reordering extension. For the German→English (Section 4.3) and English→German (Section 4.4) tasks, we utilize morpho-syntactic analysis to preprocess the data (Section 4.1) and employ system combination to produce a consensus hypothesis from normal and reverse translations (Section 4.2) of phrase-based and hierarchical phrase-based setups.

[1] http://www.statmt.org/wmt12/translation-task.html

2 Translation Systems

2.1 Phrase-Based System

The phrase-based translation (PBT) system used in this work is an in-house implementation of the state-of-the-art decoder described in (Zens and Ney, 2008). We use the standard set of models with phrase translation probabilities and lexical smoothing in both directions, word and phrase penalty, distance-based distortion model, an n-gram target language model and three binary count features. The parameter weights are optimized with minimum error rate training (MERT) (Och, 2003).

2.2 Hierarchical Phrase-Based System

For our hierarchical phrase-based translation (HPBT) setups, we employ the open source translation toolkit Jane (Vilar et al., 2010; Stein et al., 2011; Vilar et al., 2012), which has been developed at RWTH and is freely available for non-commercial use. In hierarchical phrase-based translation (Chiang, 2007), a weighted synchronous context-free grammar is induced from parallel text.

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 304–311,
Montréal, Canada, June 7-8, 2012.

In addition to contiguous *lexical* phrases, *hierarchical* phrases with up to two gaps are extracted. The search is carried out with a parsing-based procedure. The standard models integrated into our Jane systems are: phrase translation probabilities and lexical smoothing probabilities in both translation directions, word and phrase penalty, binary features marking hierarchical phrases, glue rule, and rules with non-terminals at the boundaries, four binary count features, and an n-gram language model. Optional additional models comprise IBM model 1 (Brown et al., 1993), discriminative word lexicon (DWL) models and triplet lexicon models (Mauser et al., 2009), discriminative reordering extensions (Huck et al., 2011a), insertion and deletion models (Huck and Ney, 2012), and several syntactic enhancements like preference grammars (Stein et al., 2010) and string-to-dependency features (Peter et al., 2011). We utilize the cube pruning algorithm (Huang and Chiang, 2007) for decoding and optimize the model weights with MERT.

2.3 System Combination

System combination is used to produce consensus translations from multiple hypotheses generated with different translation engines. The basic concept of RWTH's approach to machine translation system combination is described in (Matusov et al., 2006; Matusov et al., 2008). This approach includes an enhanced alignment and reordering framework. A lattice is built from the input hypotheses. The translation with the best score within the lattice according to a couple of statistical models is selected as consensus translation.

2.4 Other Tools and Techniques

We employ GIZA++ (Och and Ney, 2003) to train word alignments. The two trained alignments are heuristically merged to obtain a symmetrized word alignment for phrase extraction. All language models (LMs) are created with the SRILM toolkit (Stolcke, 2002) and are standard 4-gram LMs with interpolated modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998). We evaluate in truecase, using the BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) measures.

		French	English
EP + NC	Sentences	2.1M	
	Running Words	63.3M	57.6M
	Vocabulary	147.8K	128.5K
	Singletons	5.4K	5.1K
+ 10^9	Sentences	22.9M	
	Running Words	728.6M	624.0M
	Vocabulary	1.7M	1.7M
	Singletons	0.8M	0.8M
+ UN	Sentences	35.4M	
	Running Words	1 113.5M	956.4M
	Vocabulary	1.9M	2.0M
	Singletons	0.9M	1.0M

Table 1: Corpus statistics of the preprocessed French-English parallel training data. *EP* denotes Europarl, *NC* denotes News Commentary. In the data, numerical quantities have been replaced by a single category symbol.

3 French-English Setups

We trained phrase-based translation systems for French→English and hierarchical phrase-based translation systems for English→French. Corpus statistics for the French-English parallel data are given in Table 1. The LMs are 4-grams trained on the provided resources for the respective language (Europarl, News Commentary, UN, 10^9, and monolingual News Crawl language model training data).[2] For French→English we also investigate a smaller English LM on Europarl and News Commentary data only. For English→French we experiment with additional target-side data from the LDC French Gigaword Second Edition (LDC2009T28), which is an archive of newswire text data that has been acquired over several years by the LDC.[3] The LDC French Gigaword v2 is permitted for constrained submissions in the WMT shared translation task. As a development set for MERT, we use newstest2009 in all setups.

3.1 Experimental Results French→English

For the French→English task, the phrase-based SMT system (PBT) is set up using the standard models listed in Section 2.1. We vary the training data we use to train the system and compare the results.

[2]The parallel 10^9 corpus is often also referred to as *WMT Giga French-English release 2*.

[3]http://www.ldc.upenn.edu

French→English	newstest2008		newstest2009		newstest2010		newstest2011	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
PBT baseline	20.3	63.8	23.0	60.0	23.2	59.1	24.7	57.3
+ LM: +10^9+UN	22.5	61.4	26.2	57.3	26.6	56.1	27.7	54.5
+ TM: +10^9	23.3	60.8	27.6	56.2	27.6	55.4	29.1	53.4

Table 2: Results for the French→English task (truecase). newstest2009 is used as development set. BLEU and TER are given in percentage.

English→French	newstest2008		newstest2009		newstest2010		newstest2011	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
HPBT	20.9	66.0	23.6	62.5	25.1	60.2	27.4	57.6
+ 10^9 and UN	22.5	63.2	25.4	59.8	27.0	57.1	29.9	53.9
+ LDC Gigaword v2	23.0	63.0	25.9	59.4	27.3	56.9	29.6	54.1
+ insertion model	23.0	62.9	26.1	59.2	27.2	56.8	30.0	53.7
+ noisy-or lexical scores	23.2	62.5	26.1	59.0	27.6	56.4	30.2	53.4
+ DWL	23.3	62.5	26.2	58.9	27.9	55.9	30.4	53.2
+ IBM-1	23.4	62.3	26.2	58.8	28.0	55.7	30.4	53.1
+ discrim. RO	23.5	62.2	26.7	58.5	28.1	55.9	30.8	52.8

Table 3: Results for the English→French task (truecase). newstest2009 is used as development set. BLEU and TER are given in percentage.

It should be noted that these setups do not use any English LDC Gigaword data for LM training at all.

Our baseline system uses the Europarl and News Commentary data for training LM and phrase table. Corpus statistics are shown in the "EP+NC" section of Table 1. This results in a performance of 24.7 points BLEU on newstest2011. Then we add the 10^9 as well as UN data and more monolingual English data from the News Crawl corpus to the data used for training the language model. This system obtains a score of 27.7 points BLEU on newstest2011. Our final system uses Europarl, News Commentary, 10^9 and UN data and News Crawl monolingual data for LM training and the Europarl, News Commentary and 10^9 data (Table 1) for phrase table training. Using these data sets the system reaches 29.1 points BLEU.

The experimental results are summarized in Table 2.

3.2 Experimental Results English→French

For the English→French task, the baseline system is a hierarchical phrase-based setup including the standard models as listed in Section 2.2, apart from the binary count features. We limit the recursion depth for hierarchical rules with a shallow-1 grammar (de Gispert et al., 2010).

In a shallow-1 grammar, the generic non-terminal X of the standard hierarchical approach is replaced by two distinct non-terminals XH and XP. By changing the left-hand sides of the rules, lexical phrases are allowed to be derived from XP only, hierarchical phrases from XH only. On all right-hand sides of hierarchical rules, the X is replaced by XP. Gaps within hierarchical phrases can thus solely be filled with purely lexicalized phrases, but not a second time with hierarchical phrases. The initial rule is substituted with

$$\begin{aligned} S &\rightarrow \langle XP^{\sim 0}, XP^{\sim 0} \rangle \\ S &\rightarrow \langle XH^{\sim 0}, XH^{\sim 0} \rangle , \end{aligned} \tag{1}$$

and the glue rule is substituted with

$$\begin{aligned} S &\rightarrow \langle S^{\sim 0} XP^{\sim 1}, S^{\sim 0} XP^{\sim 1} \rangle \\ S &\rightarrow \langle S^{\sim 0} XH^{\sim 1}, S^{\sim 0} XH^{\sim 1} \rangle . \end{aligned} \tag{2}$$

The main benefit of a restriction of the recursion depth is a gain in decoding efficiency, thus allowing us to set up systems more rapidly and to explore more model combinations and more system configurations.

The experimental results for English→French are given in Table 3. Starting from the shallow hierarchical baseline setup on Europarl and News Commentary parallel data only (but Europarl, News Commentary, 10^9, UN, and News Crawl data for LM training), we are able to improve translation quality considerably by first adopting more parallel (10^9 and UN) and monolingual (French LDC Gigaword v2) training resources and then employing several different models that are not included in the baseline already. We proceed with individual descriptions of the methods we use and report their respective effect in BLEU on the test sets.

10^9 and UN (up to +2.5 points BLEU) While the amount of provided parallel data from Europarl and News Commentary sources is rather limited (around 2M sentence pairs in total), the UN and the 10^9 corpus each provide a substantial collection of further training material. By appending both corpora, we end up at roughly 35M parallel sentences (cf. Table 1). We utilize this full amount of data in our system, but extract a phrase table with only lexical (i.e. non-hierarchical) phrases from the full parallel data. We add it as a second phrase table to the baseline system, with a binary feature that enables the system to reward or penalize the application of phrases from this table.

LDC Gigaword v2 (up to +0.5 points BLEU) The LDC French Gigaword Second Edition (LDC2009T28) provides some more monolingual French resources. We include a total of 28.2M sentences from both the AFP and APW collections in our LM training data.

insertion model (up to +0.4 points BLEU) We add an insertion model to the log-linear model combination. This model is designed as a means to avoid the omission of content words in the hypotheses. It is implemented as a phrase-level feature function which counts the number of inserted words. We apply the model in source-to-target and target-to-source direction. A target-side word is considered inserted based on lexical probabilities with the words on the foreign language side of the phrase, and vice versa for a source-side word. As thresholds, we compute individual arithmetic averages for each word from the vocabulary (Huck and Ney, 2012).

noisy-or lexical scores (up to +0.4 points BLEU) In our baseline system, the $t_{\text{Norm}}(\cdot)$ lexical scoring variant as described in (Huck et al., 2011a) is employed with a relative frequency (RF) lexicon model for phrase table smoothing. The single-word based translation probabilities of the RF lexicon model are extracted from word-aligned parallel training data, in the fashion of (Koehn et al., 2003). We exchange the baseline lexical scoring with a noisy-or (Zens and Ney, 2004) lexical scoring variant $t_{\text{NoisyOr}}(\cdot)$.

DWL (up to +0.3 points BLEU) We augment our system with phrase-level lexical scores from discriminative word lexicon (DWL) models (Mauser et al., 2009; Huck et al., 2011a) in both source-to-target and target-to-source direction. The DWLs are trained on News Commentary data only.

IBM-1 (up to +0.1 points BLEU) On News Commentary and Europarl data, we train IBM model-1 (Brown et al., 1993) lexicons in both translation directions and also use them to compute phrase-level scores.

discrim. RO (up to +0.4 points BLEU) The modification of the grammar to a shallow-1 version restricts the search space of the decoder and is convenient to prevent overgeneration. In order not to be too restrictive, we reintroduce more flexibility into the search process by extending the grammar with specific reordering rules

$$\begin{aligned} XP &\rightarrow \langle XP^{\sim 0} XP^{\sim 1}, XP^{\sim 1} XP^{\sim 0} \rangle \\ XP &\rightarrow \langle XP^{\sim 0} XP^{\sim 1}, XP^{\sim 0} XP^{\sim 1} \rangle \,. \end{aligned} \tag{3}$$

The upper rule in Equation (3) is a swap rule that allows adjacent lexical phrases to be transposed, the lower rule is added for symmetry reasons, in particular because sequences assembled with these rules are allowed to fill gaps within hierarchical phrases. Note that we apply a length constraint of 10 to the number of terminals spanned by an XP. We introduce two binary indicator features, one for each of the two rules in Equation (3). In addition to adding

	German	English
Sentences	2.0M	
Running Words	55.3M	55.7M
Vocabulary	191.6K	129.0K
Singletons	75.5K	51.8K

Table 4: Corpus statistics of the preprocessed German-English parallel training data (Europarl and News Commentary). In the data, numerical quantities have been replaced by a single category symbol.

these rules, a discriminatively trained lexicalized reordering model is applied (Huck et al., 2012).

4 German-English Setups

We trained phrase-based and hierarchical translation systems for both translation directions of the German-English language pair. Corpus statistics for German-English can be found in Table 4. The language models are 4-grams trained on the respective target side of the bilingual data as well as on the provided News Crawl corpus. For the English language model the 10^9 French-English, UN and LDC Gigaword Fourth Edition corpora are used additionally. For the 10^9 French-English, UN and LDC Gigaword corpora we apply the data selection technique described in (Moore and Lewis, 2010). We examine two different language models, one with LDC data and one without. All German→English systems are optimized on newstest2010. For English→German, we use newstest2009 as development set. The newstest2011 set is used as test set and the scores for newstest2008 are included for completeness.

4.1 Morpho-Syntactic Analysis

In order to reduce the source vocabulary size for the German→English translation, the German text is preprocessed by splitting German compound words with the frequency-based method described in (Koehn and Knight, 2003). To further reduce translation complexity of PBT, we employ the long-range part-of-speech based reordering rules proposed by Popović and Ney (2006).

4.2 Reverse Translation

For reverse translations we need to change the word order of the bilingual corpus. For example, if we reverse both source and target language, the original training example "der Hund mag die Katze . → the dog likes the cat ." is converted into a new training example ". Katze die mag Hund der → . cat the likes dog the". We call this type of modification of source or target language *reversion*. A system trained of this data is called *reverse*. This modification changes the corpora and hence the language model and alignment training produce different results.

4.3 Experimental Results German→English

Our results for the German→English task are shown in Table 5. For this task, we apply the idea of reverse translation for both the phrase-based and the hierarchical approach. It seems that the reversed systems perform slightly worse. However, when we employ system combination using both reverse translation setups (*PBT reverse* and *HPBT reverse*) and both baseline setups (*PBT baseline* and *HPBT baseline*), the translation quality is improved by up to 0.4 points in BLEU and 1.0 points TER compared to the best single system.

The addition of LDC Gigaword corpora (+GW) to the language model training data of the baseline setups shows improvements in both BLEU and TER. Furthermore, with the system combination including these setups, we are able to report an improvement of up to 0.7 points BLEU and 1.0 points TER over the best single setup. Compared to the system combination based on systems which are not using the LDC Gigaword corpora, we gain 0.3 points in BLEU and 0.4 points in TER.

4.4 Experimental Results English→German

Our results for the English→German task are shown in Table 6. For this task, we first compare systems using one, two or three language models of different parts of the data. The language model for systems with only one language model is created with all monolingual and parallel data. A language model with all monolingual data and a language model with all parallel data is created for the systems with two language models. For the systems with three language models, we also split the parallel data in two parts consisting of either only Europarl data or only News Commentary data. For PBT the system with two language models performs best for all test sets. Further, we apply the idea of reverse

German→English	newstest2008		newstest2009		newstest2010		newstest2011	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
PBT baseline	21.1	62.3	20.8	61.4	23.7	59.3	21.3	61.3
PBT reverse	20.8	62.4	20.6	61.5	23.6	59.2	21.2	61.2
HPBT baseline	21.3	62.5	20.9	61.7	23.9	59.4	21.3	61.6
HPBT reverse	21.2	63.5	20.9	62.0	23.6	59.2	21.4	61.9
system combination (secondary)	21.5	61.6	21.2	60.6	24.3	58.3	21.7	60.3
PBT baseline +GW	21.5	61.9	21.2	61.1	24.0	59.0	21.3	61.4
PBT reverse	20.8	62.4	20.6	61.5	23.6	59.2	21.2	61.2
HPBT baseline +GW	21.6	62.3	21.3	61.3	24.0	59.4	21.6	61.5
HPBT reverse	21.2	63.5	20.9	62.0	23.6	59.2	21.4	61.9
system combination (primary)	21.9	61.2	21.4	60.5	24.7	58.0	21.9	60.2

Table 5: Results for the German→English task (truecase). *+GW* denotes the usage of LDC Gigaword data for the language model, newstest2010 serves as development set. BLEU and TER are given in percentage.

English→German	newstest2008		newstest2009		newstest2010		newstest2011	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
PBT baseline 1 LM	14.6	71.7	14.8	70.8	15.8	66.9	15.3	70.0
PBT baseline 2 LM (*)	14.9	70.9	14.9	70.4	16.0	66.3	15.4	69.5
PBT baseline 3 LM	14.8	71.5	14.9	70.5	16.0	66.7	15.1	70.1
PBT reverse 2 LM (*)	14.9	71.4	15.1	70.2	15.9	66.5	15.0	69.7
HPBT baseline 2 LM (*)	15.1	71.8	15.3	71.1	16.2	67.4	15.4	70.3
HPBT baseline 2 LM opt on 4bleu-ter	15.2	68.4	15.0	67.7	15.9	64.6	15.1	67.1
HPBT reverse 2 LM (*)	15.4	71.3	15.3	70.7	16.7	66.9	15.5	70.1
syscombi of (*)	15.6	69.2	15.4	68.9	16.5	65.0	15.6	68.0

Table 6: Results for the English→German task (truecase). newstest2009 is used as development set. BLEU and TER are given in percentage.

translation for both the phrase-based and the hierarchical approach. The *PBT reverse 2 LM* systems perform slightly worse compared to *PBT baseline 2 LM*. The *HPBT reverse 2 LM* performs better compared to *HPBT baseline 2 LM*. When we employ system combination using both reverse translation setups (*PBT reverse 2 LM* and *HPBT reverse 2 LM*) and both baseline setups (*PBT baseline 2 LM* and *HPBT baseline 2 LM*), the translation quality is improved by up to 0.2 points in BLEU and 2.1 points in TER compared to the best single system.

5 Conclusion

For the participation in the WMT 2012 shared translation task, RWTH experimented with both phrase-based and hierarchical translation systems. Several different techniques were evaluated and yielded considerable improvements over the respective baseline systems as well as over our last year's setups (Huck et al., 2011b). Among these techniques are an insertion model, the noisy-or lexical scoring variant, additional phrase-level lexical scores from IBM model 1 and discriminative word lexicon models, a discriminative reordering extension for hierarchical translation, reverse translation, and system combination.

Acknowledgments

This work was achieved as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathemat-

ics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, June.

Stanley F. Chen and Joshua Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, Massachusetts, USA, August.

David Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228.

Adrià de Gispert, Gonzalo Iglesias, Graeme Blackwood, Eduardo R. Banga, and William Byrne. 2010. Hierarchical Phrase-Based Translation with Weighted Finite-State Transducers and Shallow-n Grammars. *Computational Linguistics*, 36(3):505–533.

Liang Huang and David Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 144–151, Prague, Czech Republic, June.

Matthias Huck and Hermann Ney. 2012. Insertion and Deletion Models for Statistical Machine Translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies conference*, Montreal, Canada, June.

Matthias Huck, Saab Mansour, Simon Wiesler, and Hermann Ney. 2011a. Lexicon Models for Hierarchical Phrase-Based Machine Translation. In *International Workshop on Spoken Language Translation*, pages 191–198, San Francisco, California, USA, December.

Matthias Huck, Joern Wuebker, Christoph Schmidt, Markus Freitag, Stephan Peitz, Daniel Stein, Arnaud Dagnelies, Saab Mansour, Gregor Leusch, and Hermann Ney. 2011b. The RWTH Aachen Machine Translation System for WMT 2011. In *EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 405–412, Edinburgh, UK, July.

Matthias Huck, Stephan Peitz, Markus Freitag, and Hermann Ney. 2012. Discriminative Reordering Extensions for Hierarchical Phrase-Based Machine Translation. In *16th Annual Conference of the European Association for Machine Translation*, Trento, Italy, May.

Reinhard Kneser and Hermann Ney. 1995. Improved Backing-Off for M-gram Language Modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, May.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of European Chapter of the ACL (EACL 2009)*, pages 187–194.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. of the Human Language Technology Conf. (HLT-NAACL)*, pages 127–133, Edmonton, Canada, May/June.

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 33–40, Trento, Italy, April.

E. Matusov, G. Leusch, R.E. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y.-S. Lee, J.B. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney. 2008. System Combination for Machine Translation of Spoken and Written Language. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1222–1237.

Arne Mauser, Saša Hasan, and Hermann Ney. 2009. Extending Statistical Machine Translation with Discriminative and Trigger-Based Lexicon Models. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 210–218, Singapore, August.

Robert C. Moore and William Lewis. 2010. Intelligent Selection of Language Model Training Data. In *ACL (Short Papers)*, pages 220–224, Uppsala, Sweden, July.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Jan-Thorsten Peter, Matthias Huck, Hermann Ney, and Daniel Stein. 2011. Soft String-to-Dependency Hierarchical Machine Translation. In *International Workshop on Spoken Language Translation*, pages 246–253, San Francisco, California, USA, December.

Maja Popović and Hermann Ney. 2006. POS-based Word Reorderings for Statistical Machine Translation. In *International Conference on Language Resources and Evaluation*, pages 1278–1283, Genoa, Italy, May.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 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, August.

Daniel Stein, Stephan Peitz, David Vilar, and Hermann Ney. 2010. A Cocktail of Deep Syntactic Features for Hierarchical Machine Translation. In *Conf. of the Association for Machine Translation in the Americas (AMTA)*, Denver, Colorado, USA, October/November.

Daniel Stein, David Vilar, Stephan Peitz, Markus Freitag, Matthias Huck, and Hermann Ney. 2011. A Guide to Jane, an Open Source Hierarchical Translation Toolkit. *The Prague Bulletin of Mathematical Linguistics*, (95):5–18, April.

Andreas Stolcke. 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, Colorado, USA, September.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 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, July.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2012. Jane: an advanced freely available hierarchical machine translation toolkit. *Machine Translation*, pages 1–20. http://dx.doi.org/10.1007/s10590-011-9120-y.

Richard Zens and Hermann Ney. 2004. Improvements in Phrase-Based Statistical Machine Translation. In *Proc. Human Language Technology Conf. / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL)*, pages 257–264, Boston, Massachusetts, USA, May.

Richard Zens and Hermann Ney. 2008. Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation. In *International Workshop on Spoken Language Translation*, pages 195–205, Honolulu, Hawaii, USA, October.