The RWTH Aachen Machine Translation System for WMT 2013

Stephan Peitz, Saab Mansour, Jan-Thorsten Peter, Christoph Schmidt, Joern Wuebker, Matthias Huck, Markus Freitag 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 *ACL 2013 Eighth Workshop on Statistical Machine Translation* (WMT 2013). 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 hierarchical phrase reordering, translation model interpolation, domain adaptation techniques, weighted phrase extraction, word class language model, continuous space language model and system combination. By application of these methods we achieve considerable improvements over the respective baseline systems.

1 Introduction

For the WMT 2013 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 3.2), we apply a standard phrase-based system with up to five language models including a word class language model. In addition, we employ translation model interpolation and hierarchical phrase reordering. For the English→French task (Section 3.1), we train translation models on different training data sets and augment the phrase-based system with a hierarchical reordering model, a word class language model, a discriminative word lexicon and a insertion and deletion model. 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), domain-adaptation (Section 4.2) and a hierarchical reordering model. For the German→English task, an augmented hierarchical phrase-based system is set up and we rescore the phrase-based baseline with a continuous space language model. Finally, we perform a system combination.

2 Translation Systems

In this evaluation, we employ phrase-based translation and hierarchical phrase-based translation. Both approaches are implemented in *Jane* (Vilar et al., 2012; Wuebker et al., 2012), a statistical machine translation toolkit which has been developed at RWTH Aachen University and is freely available for non-commercial use.[2]

2.1 Phrase-based System

In the phrase-based decoder (source cardinality synchronous search, *SCSS*), 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. Optional additional models used in this evaluation are the hierarchical reordering model (*HRM*) (Galley and Manning, 2008), a word class language model (*WCLM*) (Wuebker et

[1] http://www.statmt.org/wmt13/translation-task.html

[2] http://www.hltpr.rwth-aachen.de/jane/

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 193–199,
Sofia, Bulgaria, August 8-9, 2013

al., 2012), a discriminative word lexicon (*DWL*) (Mauser et al., 2009), and insertion and deletion models (*IDM*) (Huck and Ney, 2012). The parameter weights are optimized with minimum error rate training (MERT) (Och, 2003). The optimization criterion is BLEU.

2.2 Hierarchical Phrase-based System

In hierarchical phrase-based translation (Chiang, 2007), a weighted synchronous context-free grammar is induced from parallel text. In addition to continuous *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 hierarchical systems (Vilar et al., 2010; Huck et al., 2012c) 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 and triplet lexicon models (Mauser et al., 2009; Huck et al., 2011), discriminative reordering extensions (Huck et al., 2012a), insertion and deletion models (Huck and Ney, 2012), and several syntactic enhancements like preference grammars (Stein et al., 2010) and soft string-to-dependency features (Peter et al., 2011). We utilize the cube pruning algorithm for decoding (Huck et al., 2013) and optimize the model weights with MERT. The optimization criterion is BLEU.

2.3 System Combination

System combination is used to produce consensus translations from multiple hypotheses generated with different translation engines. First, a word to word alignment for the given single system hypotheses is produced. In a second step a confusion network is constructed. Then, the hypothesis with the highest probability is extracted from this confusion network. For the alignment procedure, one of the given single system hypotheses is chosen as primary system. To this primary system all other hypotheses are aligned using the METEOR (Lavie and Agarwal, 2007) alignment and thus the primary system defines the word order. Once the alignment is given, the corresponding confusion network is constructed. An example is given in Figure 1.

The model weights of the system combination are optimized with standard MERT on 100-best lists. For each single system, a factor is added to the log-linear framework of the system combination. Moreover, this log-linear model includes a word penalty, a language model trained on the input hypotheses, a binary feature which penalizes word deletions in the confusion network and a primary feature which marks the system which provides the word order. The optimization criterion is 4BLEU-TER.

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). The Stanford Parser (Klein and Manning, 2003) is used to obtain parses of the training data for the syntactic extensions of the hierarchical system. We evaluate in truecase with BLEU (Papineni et al., 2002) and TER (Snover et al., 2006).

2.5 Filtering of the Common Crawl Corpus

The new Common Crawl corpora contain a large number of sentences that are not in the labelled language. To clean these corpora, we first extracted a vocabulary from the other provided corpora. Then, only sentences containing at least 70% word from the known vocabulary were kept. In addition, we discarded sentences that contain more words from target vocabulary than source vocabulary on the source side. These heuristics reduced the French-English Common Crawl corpus by 5,1%. This filtering technique was also applied on the German-English version of the Common Crawl corpus.

3 French–English Setups

We trained phrase-based translation systems for French→English and 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, Common Crawl, and monolingual News Crawl

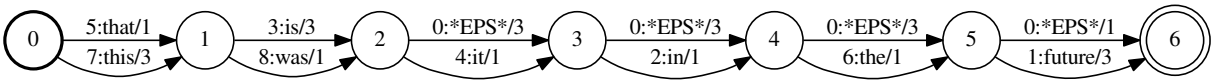

Figure 1: Confusion network of four different hypotheses.

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

		French	English
EPPS + NC	Sentences	2.2M	
	Running Words	64.7M	59.7M
	Vocabulary	153.4K	132.2K
CC	Sentences	3.2M	
	Running Words	88.1M	80.9.0M
	Vocabulary	954.8K	908.0K
UN	Sentences	12.9M	
	Running Words	413.3M	362.3M
	Vocabulary	487.1K	508.3K
10^9	Sentences	22.5M	
	Running Words	771.7M	661.1M
	Vocabulary	1 974.0K	1 947.2K
All	Sentences	40.8M	
	Running Words	1 337.7M	1 163.9M
	Vocabulary	2 749.8K	2 730.1K

language model training data).[3]

3.1 Experimental Results English→French

For the English→French task, separate translation models (TMs) were trained for each of the five data sets and fed to the decoder. Four additional indicator features are introduced to distinguish the different TMs. Further, we applied the hierarchical reordering model, the word class language model, the discriminative word lexicon, and the insertion and deletion model. Table 2 shows the results of our experiments.

As a development set for MERT, we use newstest2010 in all setups.

3.2 Experimental Results French→English

For the French→English task, a translation model (*TM*) was trained on all available parallel data. For the baseline, we interpolated this TM with an in-domain TM trained on EPPS+NC and employed the hierarchical reordering model. Moreover, three language models were used: The first language model was trained on the English side of all available parallel data, the second one on EPPS and NC and the third LM on the News Shuffled data. The baseline was improved by adding a fourth LM trained on the Gigaword corpus (Version 5) and a 5-gram word class language model trained on News Shuffled data. For the WCLM, we used 50 word classes clustered with the tool *mkcls* (Och, 2000). All results are presented in Table 3.

4 German–English Setups

For both translation directions of the German-English language pair, we trained phrase-based translation systems. 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 Fifth Edition corpora are used additionally.

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, we employ the long-range part-of-speech based reordering rules proposed by Popović and Ney (2006).

4.2 Domain Adaptation

This year, we experimented with filtering and weighting for domain-adaptation for the German-English task. To perform adaptation, we define a general-domain (GD) corpus composed from the news-commentary, europarl and Common Crawl corpora, and an in-domain (ID) corpus using a concatenation of the test sets (newstest{2008, 2009, 2010, 2011, 2012}) with the corresponding references. We use the test sets as in-domain

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

Table 2: Results for the English→French task (truecase). newstest2010 is used as development set. BLEU and TER are given in percentage.

	newstest2008		newstest2009		newstest2010		newstest2011		newstest2012	
English→French	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
TM:EPPS + HRM	22.9	63.0	25.0	60.0	27.8	56.7	28.9	54.4	27.2	57.1
TM:UN + HRM	22.7	63.4	25.0	60.0	28.3	56.4	29.5	54.2	27.3	57.1
TM:10^9 + HRM	23.5	62.3	26.0	59.2	29.6	55.2	30.3	53.3	28.0	56.4
TM:CC + HRM	23.5	62.3	26.2	58.8	29.2	55.3	30.3	53.3	28.2	56.0
TM:NC	21.0	64.8	22.3	61.6	25.6	58.7	26.9	56.6	25.7	58.5
+ HRM	21.5	64.3	22.6	61.2	26.1	58.4	27.3	56.1	26.0	58.2
+ TM:EPPS,CC,UN	23.9	61.8	26.4	58.6	29.9	54.7	31.0	52.7	28.6	55.6
+ TM:10^9	24.0	61.5	26.5	58.4	30.2	54.2	31.1	52.3	28.7	55.3
+ WCLM, DWL, IDM	24.0	61.6	26.5	58.3	30.4	54.0	31.4	52.1	28.8	55.2

Table 3: Results for the French→English task (truecase). newstest2010 is used as development set. BLEU and TER are given in percentage.

	newstest2010		newstest2011		newstest2012	
French→English	BLEU	TER	BLEU	TER	BLEU	TER
SCSS baseline	28.1	54.6	29.1	53.3	-	-
+ GigaWord.v5 LM	28.6	54.2	29.6	52.9	29.6	53.3
+ WCLM	29.1	53.8	30.1	52.5	29.8	53.1

(newswire) as the other corpora are coming from differing domains (news commentary, parliamentary discussions and various web sources), and on initial experiments, the other corpora did not perform well when used as an in-domain representative for adaptation. To check whether over-fitting occurs, we measure the results of the adapted systems on the evaluation set of this year (newstest2013) which was not used as part of the in-domain set.

The filtering experiments are done similarly to (Mansour et al., 2011), where we compare filtering using LM and a combined LM and IBM Model 1 (LM+M1) based scores. The scores for each sentence pair in the general-domain corpus are based on the bilingual cross-entropy difference of the in-domain and general-domain models. Denoting $H_{LM}(x)$ as the cross entropy of sentence x according to LM, then the cross entropy difference $DH_{LM}(x)$ can be written as:

$$DH_{LM}(x) = H_{LM_{ID}}(x) - H_{LM_{GD}}(x)$$

The bilingual cross entropy difference for a sentence pair (s, t) in the GD corpus is then defined by:

$$DH_{LM}(s) + DH_{LM}(t)$$

For IBM Model 1 (M1), the cross-entropy $H_{M1}(s|t)$ is defined similarly to the LM cross-entropy, and the resulting bilingual cross-entropy difference will be of the form:

$$DH_{M1}(s|t) + DH_{M1}(t|s)$$

The combined LM+M1 score is obtained by summing the LM and M1 bilingual cross-entropy difference scores. To perform filtering, the GD corpus sentence pairs are scored by the appropriate method, sorted by the score, and the n-best sentences are then used to build an adapted system.

In addition to adaptation using filtering, we experiment with weighted phrase extraction similar to (Mansour and Ney, 2012). We differ from their work by using a combined LM+M1 weight to perform the phrase extraction instead of an LM based weight. We use a combined LM+M1 weight as this worked best in the filtering experiments, making scoring with LM+M1 more reliable than LM scores only.

4.3 Experimental Results German→English

For the German→English task, the baseline is trained on all available parallel data and includes the hierarchical reordering model. The results of the various filtering and weighting experiments are summarized in Table 5.

Table 5: German-English results (truecase). BLEU and TER are given in percentage. Corresponding development set is marked with *. † labels the single systems selected for the system combination.

	newstest2009		newstest2010		newstest2011		newstest2012		newstest2013	
German→English	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
SCSS baseline	21.7	61.1	24.8*	58.9*	22.0	61.1	23.4	60.0	26.1	56.4
LM 800K-best	21.6	60.5	24.7*	58.3*	22.0	60.5	23.6	59.7	-	-
LM+M1 800K-best	21.4	60.5	24.7*	58.1*	22.0	60.4	23.7	59.2	-	-
(LM+M1)*TM	22.1	60.2	25.4*	57.8*	22.5	60.1	24.0	59.1	-	-
(LM+M1)*TM+GW	22.8	59.5	25.7*	57.2*	23.1	59.5	24.4	58.6	26.6	55.5
(LM+M1)*TM+GW†	22.9*	61.1*	25.2	59.3	22.8	61.5	23.7	60.8	26.4	57.1
SCSS baseline	22.6*	61.6*	24.1	60.1	22.1	62.0	23.1	61.2	-	-
CSLM rescoring†	22.0	60.4	25.1*	58.3*	22.4	60.2	23.9	59.3	26.0	56.0
HPBT†	21.9	60.4	24.9*	58.2*	22.3	60.3	23.6	59.6	25.9	56.3
system combination	-	-	-	-	23.4*	59.3*	24.7	58.5	27.1	55.3

Table 6: English-German results (truecase). newstest2009 was used as development set. BLEU and TER are given in percentage.

	newstest2008		newstest2009		newstest2010		newstest2011		newstest2012	
English→German	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
SCSS baseline	14.9	70.9	14.9	70.4	16.0	66.3	15.4	69.5	15.7	67.5
LM 800K-best	15.1	70.9	15.1	70.3	16.2	66.3	15.6	69.4	15.9	67.4
(LM+M1) 800K-best	15.8	70.8	15.4	70.0	16.2	66.2	16.0	69.3	16.1	67.4
(LM+M1) ifelse	16.1	70.6	15.7	69.9	16.5	66.0	16.2	69.2	16.3	67.2

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

	German	English
Sentences	4.1M	
Running Words	104M	104M
Vocabulary	717K	750K

For filtering, we use the 800K best sentences from the whole training corpora, as this selection performed best on the dev set among 100K,200K,400K,800K,1600K setups. Filtering seems to mainly improve on the TER scores, BLEU scores are virtually unchanged in comparison to the baseline. LM+M1 filtering improves further on TER in comparison to LM-based filtering.

The weighted phrase extraction performs best in our experiments, where the weights from the LM+M1 scoring method are used. Improvements in both BLEU and TER are achieved, with BLEU improvements ranging from +0.4% up-to +0.6% and TER improvements from -0.9% and up-to -1.1%.

As a final step, we added the English Gigaword corpus to the LM (+GW). This resulted in further improvements of the systems.

In addition, the system as described above was tuned on newstest2009. Using this development set results in worse translation quality.

Furthermore, we rescored the SCSS baseline tuned on newstest2009 with a continuous space language model (*CSLM*) as described in (Schwenk et al., 2012). The CSLM was trained on the europarl and news-commentary corpora. For rescoring, we used the newstest2011 set as tuning set and re-optimized the parameters with MERT on 1000-best lists. This results in an improvement of up to 0.8 points in BLEU compared to the baseline.

We compared the phrase-based setups with a hierarchical translation system, which was augmented with preference grammars, soft string-to-dependency features, discriminative reordering extensions, DWL, IDM, and discriminative re-

ordering extensions. The phrase table of the hierarchical setup has been extracted from News Commentary and Europarl parallel data only (not from Common Crawl).

Finally, three setups were joined in a system combination and we gained an improvement of up to 0.5 points in BLEU compared to the best single system.

4.4 Experimental Results English→German

The results for the English→German task are shown in Table 6. While the LM-based filtering led to almost no improvement over the baseline, the LM+M1 filtering brought some improvements in BLEU. In addition to the sentence filtering, we tried to combine the translation model trained on NC+EPPS with a TM trained on Common Crawl using the *ifelse* combination (Mansour and Ney, 2012). This combination scheme concatenates both TMs and assigns the probabilities of the in-domain TM if it contains the phrase, else it uses the probabilities of the out-of-domain TM. Appling this method, we achieved further improvements.

5 Conclusion

For the participation in the WMT 2013 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., 2012b). Among these techniques are a hierarchical phrase reordering model, translation model interpolation, domain adaptation techniques, weighted phrase extraction, a word class language model, a continuous space language model 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 Mathematics 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.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 847–855, Honolulu, Hawaii, USA, October.

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*, pages 347–351, Montréal, Canada, June.

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

Matthias Huck, Stephan Peitz, Markus Freitag, and Hermann Ney. 2012a. Discriminative Reordering Extensions for Hierarchical Phrase-Based Machine Translation. In *16th Annual Conference of the European Association for Machine Translation*, pages 313–320, Trento, Italy, May.

Matthias Huck, Stephan Peitz, Markus Freitag, Malte Nuhn, and Hermann Ney. 2012b. The RWTH Aachen Machine Translation System for WMT 2012. In *NAACL 2012 Seventh Workshop on Statistical Machine Translation*, pages 304–311, Montréal, Canada, June.

Matthias Huck, Jan-Thorsten Peter, Markus Freitag, Stephan Peitz, and Hermann Ney. 2012c. Hierarchical Phrase-Based Translation with Jane 2. *The Prague Bulletin of Mathematical Linguistics*, 98:37–50, October.

Matthias Huck, David Vilar, Markus Freitag, and Hermann Ney. 2013. A Performance Study of Cube Pruning for Large-Scale Hierarchical Machine Translation. In *Proceedings of the NAACL 7th Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 29–38, Atlanta, Georgia, USA, June.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 423–430, Sapporo, Japan, July.

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.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *ACL 2007 Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June.

Saab Mansour and Hermann Ney. 2012. A Simple and Effective Weighted Phrase Extraction for Machine Translation Adaptation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 193–200, Hong Kong, December.

Saab Mansour, Joern Wuebker, and Hermann Ney. 2011. Combining Translation and Language Model Scoring for Domain-Specific Data Filtering. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 222–229, San Francisco, California, USA, December.

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.

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. 2000. mkcls: Training of word classes for language modeling. `http://www.hltpr.rwth-aachen.de/web/Software/mkcls.html`.

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.

Holger Schwenk, Anthony Rousseau, and Mohammed Attik. 2012. Large, Pruned or Continuous Space Language Models on a GPU for Statistical Machine Translation. In *NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 11–19, Montréal, Canada, June.

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.

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*, 26(3):197–216, September.

Joern Wuebker, Matthias Huck, Stephan Peitz, Malte Nuhn, Markus Freitag, Jan-Thorsten Peter, Saab Mansour, and Hermann Ney. 2012. Jane 2: Open Source Phrase-based and Hierarchical Statistical Machine Translation. In *International Conference on Computational Linguistics*, pages 483–491, Mumbai, India, December.