The RWTH Aachen Machine Translation System for WMT 2011

Matthias Huck, Joern Wuebker, Christoph Schmidt, Markus Freitag, Stephan Peitz,
Daniel Stein, Arnaud Dagnelies, Saab Mansour, Gregor Leusch and Hermann Ney
RWTH Aachen University
Aachen, Germany
surname@cs.rwth-aachen.de

Abstract

This paper describes the statistical machine translation (SMT) systems developed by RWTH Aachen University for the translation task of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation. Both phrase-based and hierarchical SMT systems were trained for the constrained German-English and French-English tasks in all directions. Experiments were conducted to compare different training data sets, training methods and optimization criteria, as well as additional models on dependency structure and phrase reordering. Further, we applied a system combination technique to create a consensus hypothesis from several different systems.

1 Overview

We sketch the baseline architecture of RWTH's setups for the WMT 2011 shared translation task by providing an overview of our translation systems in Section 2. In addition to the baseline features, we adopted several novel methods, which will be presented in Section 3. Details on the respective setups and translation results for the French-English and German-English language pairs (in both translation directions) are given in Sections 4 and 5. We finally conclude the paper in Section 6.

2 Translation Systems

For the WMT 2011 evaluation we utilized RWTH's state-of-the-art phrase-based and hierarchical translation systems as well as our in-house system combination framework. GIZA++ (Och and Ney, 2003) was employed to train word alignments, language models have been created with the SRILM toolkit (Stolcke, 2002).

2.1 Phrase-Based System

We applied a phrase-based translation (PBT) system similar to the one described in (Zens and Ney, 2008). Phrase pairs are extracted from a word-aligned bilingual corpus and their translation probability in both directions is estimated by relative frequencies. The standard feature set moreover includes an n-gram language model, phrase-level single-word lexicons and word-, phrase- and distortion-penalties. To lexicalize reordering, a discriminative reordering model (Zens and Ney, 2006a) is used. Parameters are optimized with the Downhill-Simplex algorithm (Nelder and Mead, 1965) on the word graph.

2.2 Hierarchical System

For the hierarchical setups described in this paper, the open source Jane toolkit (Vilar et al., 2010) was employed. Jane has been developed at RWTH and implements the hierarchical approach as introduced by Chiang (2007) with some state-of-the-art extensions. In hierarchical phrase-based translation, a weighted synchronous context-free grammar is induced from parallel text. In addition to contiguous *lexical* phrases, *hierarchical* phrases with up to two gaps are extracted. The search is typically carried out using the cube pruning algorithm (Huang and Chiang, 2007). The standard models integrated into our Jane systems are: phrase translation probabilities and lexical translation probabilities on phrase level, each for both translation directions, length

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 405–412,
Edinburgh, Scotland, UK, July 30–31, 2011. ©2011 Association for Computational Linguistics

penalties on word and phrase level, three binary features marking hierarchical phrases, glue rule, and rules with non-terminals at the boundaries, source-to-target and target-to-source phrase length ratios, four binary count features and an n-gram language model. The model weights are optimized with standard MERT (Och, 2003) on 100-best lists.

2.3 System Combination

System combination is used to produce consensus translations from multiple hypotheses produced with different translation engines that are better in terms of translation quality than any of the individual hypotheses. The basic concept of RWTH's approach to machine translation system combination has been described by Matusov et al. (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.

3 Translation Modeling

We incorporated several novel methods into our systems for the WMT 2011 evaluation. This section provides a short survey of three of the methods which we suppose to be of particular interest.

3.1 Language Model Data Selection

For the English and German language models, we applied the data selection method proposed in (Moore and Lewis, 2010). Each sentence is scored by the difference in cross-entropy between a language model trained from in-domain data and a language model trained from a similar-sized sample of the out-of-domain data. As in-domain data we used the news-commentary corpus. The out-of-domain data from which the data was selected are the news crawl corpus for both languages and for English the 10^9 corpus and the LDC Gigaword data. We used a 3-gram trained with the SRI toolkit to compute the cross-entropy. For the news crawl corpus, only $1/8$ of the sentences were discarded. Of the 10^9 corpus we retained $1/2$ and of the LDC Gigaword data we retained $1/4$ of the sentences to train the language models.

3.2 Phrase Model Training

For the German→English and French→English translation tasks we applied a forced alignment procedure to train the phrase translation model with the EM algorithm, similar to the one described in (DeNero et al., 2006). Here, the phrase translation probabilities are estimated from their relative frequencies in the phrase-aligned training data. The phrase alignment is produced by a modified version of the translation decoder. In addition to providing a statistically well-founded phrase model, this has the benefit of producing smaller phrase tables and thus allowing more rapid experiments. A detailed description of the training procedure is given in (Wuebker et al., 2010).

3.3 Soft String-to-Dependency

Given a dependency tree of the target language, we are able to introduce language models that span over longer distances than the usual n-grams, as in (Shen et al., 2008). To obtain dependency structures, we apply the Stanford parser (Klein and Manning, 2003) on the target side of the training material. RWTH's open source hierarchical translation toolkit Jane has been extended to include dependency information in the phrase table and to build dependency trees on the output hypotheses at decoding time from this information.

Shen et al. (2008) use only phrases that meet certain restrictions. The first possibility is what the authors call a *fixed* dependency structure. With the exception of one word within this phrase, called the *head*, no outside word may have a dependency within this phrase. Also, all inner words may only depend on each other or on the head. For a second structure, called a *floating* dependency structure, the head dependency word may also exist outside the phrase. If the dependency structure of a phrase conforms to these restrictions, it is denoted as *valid*.

In our phrase table, we mark those phrases that possess a valid dependency structure with a binary feature, but all phrases are retained as translation options. In addition to storing the dependency information, we also memorize for all hierarchical phrases if the content of gaps has been dependent on the left or on the right side. We utilize the dependency information during the search process by adding three

	French	English
Sentences	3 710 985	
Running Words	98 352 916	87 689 253
Vocabulary	179 548	216 765

Table 1: Corpus statistics of the preprocessed high-quality training data (Europarl, news-commentary, and selected parts of the 10^9 and UN corpora) for the RWTH systems for the WMT 2011 French→English and English→French translation tasks. Numerical quantities are replaced by a single category symbol.

	French	English
Sentences	29 996 228	
Running Words	916 347 538	778 544 843
Vocabulary	1 568 089	1 585 093

Table 2: Corpus statistics of the preprocessed full training data for the RWTH primary system for the WMT 2011 English→French translation task. Numerical quantities are replaced by a single category symbol.

features to the log-linear model: merging errors to the left, merging errors to the right, and the ratio of valid vs. non-valid dependency structures. The decoder computes the corresponding costs when it tries to construct a dependency tree of a (partial) hypothesis on-the-fly by merging the dependency structures of the used phrase pairs.

In an n-best reranking step, we compute dependency language model scores on the dependencies which were assembled on the hypotheses by the search procedure. We apply one language model for left-side dependencies and one for right-side dependencies. For head structures, we also compute their scores by exploiting a simple unigram language model. We furthermore include a language count feature that is incremented each time we compute a dependency language model score. As trees with few dependencies have less individual costs to be computed, they tend to obtain lower overall costs than trees with more complex structures in other sentences. The intention behind this feature is thus comparable to the word penalty in combination with a normal n-gram language model.

4 French-English Setups

We set up both hierarchical and standard phrase-based systems for the constrained condition of the WMT 2011 French→English and English→French translation tasks. The English→French RWTH primary submission was produced with a single hierarchical system, while a system combination of three systems was used to generate a final hypothesis for the French→English primary submission.

Besides the Europarl and news-commentary corpora, the provided parallel data also comprehends the large French-English 10^9 corpus and the French-English UN corpus. Since model training with such a huge amount of data requires a considerable computational effort, RWTH decided to select a high-quality part of altogether about 2 Mio. sentence pairs from the latter two corpora. The selection of parallel sentences was carried out according to three criteria: (1) Only sentences of minimum length of 4 tokens are considered, (2) at least 92% of the vocabulary of each sentence occurs in newstest2008, and (3) the ratio of the vocabulary size of a sentence and the number of its tokens is minimum 80%. Word alignments in both directions were trained with GIZA++ and symmetrized according to the refined method that was proposed in (Och and Ney, 2003). The phrase tables of the translation systems are extracted from the Europarl and news-commentary parallel training data as well as the selected high-quality parts the 10^9 and UN corpora only. The only exception is the hierarchical system used for the English→French RWTH primary submission which comprehends a second phrase table with lexical (i.e. non-hierarchical) phrases extracted from the full parallel data (approximately 30 Mio. sentence pairs).

Detailed statistics of the high-quality parallel training data (Europarl, news-commentary, and the selected parts of the 10^9 and UN corpora) are given in Table 1, the corpus statistics of the full parallel data from which the second phrase table with lexical phrases for the English→French RWTH primary system was created are presented in Table 2.

The translation systems use large 4-gram language models with modified Kneser-Ney smoothing. The French language model was trained on most of the provided French data including the monolingual LDC Gigaword corpora, the English

	newstest2009		newstest2010	
French→English	BLEU	TER	BLEU	TER
System combination of † systems (primary)	26.7	56.0	27.4	54.9
PBT with triplet lexicon, no forced alignment (contrastive) †	26.2	56.7	27.2	55.3
Jane as below + improved LM (contrastive)	26.3	57.4	26.7	56.2
Jane with parse match + syntactic labels + dependency †	26.2	57.5	26.5	56.4
PBT with forced alignment phrase training †	26.0	57.1	26.3	56.0

Table 3: RWTH systems for the WMT 2011 French→English translation task (truecase). BLEU and TER results are in percentage.

	newstest2009		newstest2010	
English→French	BLEU	TER	BLEU	TER
Jane shallow + in-domain TM + lexical phrases from full data	25.3	60.1	27.1	57.2
Jane shallow + in-domain TM + triplets + DWL + parse match	24.8	60.5	26.6	57.5
PBT with triplets, DWL, sentence-level word lexicon, discrim. reord.	24.8	60.1	26.5	57.3

Table 4: RWTH systems for the WMT 2011 English→French translation task (truecase). BLEU and TER results are in percentage.

language model was trained on automatically selected English data (cf. Section 3.1) from the provided resources including the 10^9 corpus and LDC Gigaword.

The scaling factors of the log-linear model combination are optimized towards BLEU on newstest2009, newstest2010 is used as an unseen test set.

4.1 Experimental Results French→English

The results for the French→English task are given in Table 3. RWTH's three submissions – one primary and two contrastive – are labeled accordingly in the table. The first contrastive submission is a phrase-based system with a standard feature set plus an additional triplet lexicon model (Mauser et al., 2009). The triplet lexicon model was trained on in-domain news commentary data only. The second contrastive submission is a hierarchical Jane system with three syntax-based extensions: A parse match model (Vilar et al., 2008), soft syntactic labels (Stein et al., 2010), and the soft string-to-dependency extension as described in Section 3.3. The primary submission combines the phrase-based contrastive system, a hierarchical system that is very similar to the Jane contrastive submission but with a slightly worse language model, and an additional PBT system that has been trained with forced alignment (Wuebker et al., 2010) on WMT 2010 data only.

4.2 Experimental Results English→French

The results for the English→French task are given in Table 4. We likewise submitted two contrastive systems for this translation direction. The first contrastive submission is a phrase-based system, enhanced with a triplet lexicon model and a discriminative word lexicon model (Mauser et al., 2009) – both trained on in-domain news commentary data only – as well as a sentence-level single-word lexicon model and a discriminative reordering model (Zens and Ney, 2006a). The second contrastive submission is a hierarchical Jane system with shallow rules (Iglesias et al., 2009), a triplet lexicon model, a discriminative word lexicon, the parse match model, and a second phrase table extracted from in-domain data only. Our primary submission is very similar to the latter Jane setup. It does not comprise the extended lexicon models and the parse match extension, but instead includes lexical phrases from the full 30 Mio. sentence corpus as described above.

5 German-English Setups

We trained phrase-based and hierarchical translation systems for both translation directions of the German-English language pair. The corpus statis-

	German	English
Sentences	1 857 745	
Running Words	48 449 977	50 559 217
Vocabulary	387 593	123 470

Table 5: Corpus statistics of the preprocessed training data for the WMT 2011 German→English and English→German translation tasks. Numerical quantities are replaced by a single category symbol.

tics can be found in Table 5. Word alignments were generated with GIZA++ and symmetrized as for the French-English setups.

The language models are 4-grams trained on the bilingual data as well as the provided News crawl corpus. For the English language model the 10^9 French-English and LDC Gigaword corpora were used additionally. For the 10^9 French-English and LDC Gigaword corpora RWTH applied the data selection technique described in Section 3.1. We examined two different language models, one with LDC data and one without.

Systems were optimized on the newstest2009 data set, newstest2008 was used as test set. The scores for newstest2010 are included for completeness.

5.1 Morpho-Syntactic Analysis

In order to reduce the source vocabulary size for the German→English translation, the source side was preprocessed by splitting German compound words with the frequency-based method described in (Koehn and Knight, 2003). To further reduce translation complexity, we performed the long-range part-of-speech based reordering rules proposed by (Popović et al., 2006). For additional experiments we used the TreeTagger (Schmid, 1995) to produce a lemmatized version of the German source.

5.2 Optimization Criterion

We studied the impact of different optimization criteria on tranlsation performance. The usual practice is to optimize the scaling factors to maximize BLEU. We also experimented with two different combinations of BLEU and Translation Edit Rate (TER): TER$-$BLEU and TER$-$4BLEU. The first denotes the equally weighted combination, while for the latter BLEU is weighted 4 times as strong as TER.

5.3 Experimental Results German→English

For the German→English task we conducted experiments comparing the standard phrase extraction with the phrase training technique described in Section 3.2. For the latter we applied log-linear phrase-table interpolation as proposed in (Wuebker et al., 2010). Further experiments included the use of additional language model training data, reranking of n-best lists generated by the phrase-based system, and different optimization criteria. We also carried out a system combination of several systems, including phrase-based systems on lemmatized German and on source data without compound splitting and two hierarchical systems optimized for different criteria. The results are given in Table 6.

A considerable increase in translation quality can be achieved by application of German compound splitting. The system that operates on German surface forms without compound splitting (SUR) clearly underperforms the baseline system with morphological preprocessing. The system on lemmatized German (LEM) is at about the same level as the system on surface forms.

In comparison to the standard heuristic phrase extraction technique, performing phrase training (FA) gives an improvement in BLEU on newstest2008 and newstest2009, but a degradation in TER. The addition of LDC Gigaword corpora (+GW) to the language model training data shows improvements in both BLEU and TER. Reranking was done on 1000-best lists generated by the the best available system (PBT (FA)+GW). Following models were applied: n-gram posteriors (Zens and Ney, 2006b), sentence length model, a 6-gram LM and single-word lexicon models in both normal and inverse direction. These models are combined in a log-linear fashion and the scaling factors are tuned in the same manner as the baseline system (using TER$-$4BLEU on newstest2009).

The table includes three identical Jane systems which are optimized for different criteria. The one optimized for TER$-$4BLEU offers the best balance between BLEU and TER, but was not finished in time for submission. As primary submission we chose the reranked PBT system, as secondary the system combination.

		newstest2008		newstest2009		newstest2010	
German→English	opt criterion	BLEU	TER	BLEU	TER	BLEU	TER
Syscombi of † (secondary)	TER−BLEU	21.1	62.1	20.8	61.2	23.7	59.2
Jane +GW †	BLEU	21.5	63.9	21.0	63.3	22.9	61.7
Jane +GW	TER−4BLEU	21.4	62.6	21.1	62.0	23.5	60.3
PBT (FA) rerank +GW (primary) †	TER−4BLEU	21.4	62.8	21.1	61.9	23.4	60.1
PBT (FA) +GW †	TER−4BLEU	21.1	63.0	21.1	62.2	23.3	60.3
Jane +GW †	TER−BLEU	20.9	61.1	20.4	60.5	23.4	58.3
PBT (FA)	TER−4BLEU	21.1	63.2	20.6	62.4	23.2	60.4
PBT	TER−4BLEU	20.6	62.7	20.3	61.9	23.3	59.7
PBT (SUR) †	TER−4BLEU	19.5	66.5	18.9	65.8	21.0	64.9
PBT (LEM) †	TER−4BLEU	19.2	66.1	18.9	65.4	21.0	63.5

Table 6: RWTH systems for the WMT 2011 German→English translation task (truecase). BLEU and TER results are in percentage. FA denotes systems with phrase training, +GW the use of LDC data for the language model. SUR and LEM denote the systems without compound splitting and on the lemmatized source, respectively. The three hierarchical Jane systems are identical, but used different parameter optimization criterea.

		newstest2008		newstest2009		newstest2010	
English→German	opt criterion	BLEU	TER	BLEU	TER	BLEU	TER
PBT + discrim. reord. (primary)	TER−4BLEU	15.3	70.2	15.1	69.8	16.2	65.6
PBT + discrim. reord.	BLEU	15.2	70.6	15.2	70.1	16.2	66.0
PBT	TER−4BLEU	15.2	70.7	15.2	70.2	16.2	66.1
Jane	BLEU	15.1	72.1	15.4	71.2	16.4	67.4
Jane	TER−4BLEU	15.1	68.4	14.6	69.5	14.6	65.9

Table 7: RWTH systems for the WMT 2011 English→German translation task (truecase). BLEU and TER results are in percentage.

5.4 Experimental Results English→German

We likewise studied the effect of using BLEU only versus using TER−4BLEU as optimization criterion in the English→German translation direction. Moreover, we tested the impact of the discriminative reordering model (Zens and Ney, 2006a). The results can be found in Table 7. For the phrase-based system, optimizing towards TER−4BLEU leads to slightly better results both in BLEU and TER than optimizing towards BLEU. Using the discriminative reordering model yields some improvements both on newstest2008 and newstest2010. In the case of the hierarchical system, the effect of the optimization criterion is more pronounced than for the phrase-based system. However, in this case it clearly leads to a tradeoff between BLEU and TER, as the choice of TER−4BLEU harms the translation results of test2010 with respect to BLEU.

6 Conclusion

For the participation in the WMT 2011 shared translation task, RWTH experimented with both phrase-based and hierarchical translation systems. We used all bilingual and monolingual data provided for the constrained track. To limit the size of the language model, a data selection technique was applied. Several techniques yielded improvements over the baseline, including three syntactic models, extended lexicon models, a discriminative reordering model, forced alignment training, reranking methods and different optimization criteria.

Acknowledgments

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

References

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

J. DeNero, D. Gillick, J. Zhang, and D. Klein. 2006. Why Generative Phrase Models Underperform Surface Heuristics. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 31–38.

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

G. Iglesias, A. de Gispert, E.R. Banga, and W. Byrne. 2009. Rule Filtering by Pattern for Efficient Hierarchical Translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 380–388.

D. Klein and C.D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430.

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

E. Matusov, N. Ueffing, and H. 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.

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.

A. Mauser, S. Hasan, and H. Ney. 2009. Extending Statistical Machine Translation with Discriminative and Trigger-Based Lexicon Models. In *Conference on Empirical Methods in Natural Language Processing*, pages 210–217.

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

J.A. Nelder and R. Mead. 1965. The Downhill Simplex Method. *Computer Journal*, 7:308.

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

F.J. Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proc. Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July.

M. Popović, D. Stein, and H. Ney. 2006. Statistical Machine Translation of German Compound Words. In *FinTAL - 5th International Conference on Natural Language Processing, Springer Verlag, LNCS*, pages 616–624.

H. Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland, March.

L. Shen, J. Xu, and R. Weischedel. 2008. A New String-to-Dependency Machine Translation Algorithm with a Target Dependency Language Model. In *Proceedings of ACL-08: HLT. Association for Computational Linguistics*, pages 577–585, June.

D. Stein, S. Peitz, D. Vilar, and H. Ney. 2010. A Cocktail of Deep Syntactic Features for Hierarchical Machine Translation. In *Conference of the Association for Machine Translation in the Americas 2010*, page 9, Denver, USA, October.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. Int. Conf. on Spoken Language Processing*, volume 2, pages 901 – 904, Denver, Colorado, USA, September.

D. Vilar, D. Stein, and H. Ney. 2008. Analysing Soft Syntax Features and Heuristics for Hierarchical Phrase Based Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 190–197, Waikiki, Hawaii, October.

D. Vilar, S. Stein, M. Huck, and H. 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.

J. Wuebker, A. Mauser, and H. Ney. 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, July.

R. Zens and H. Ney. 2006a. Discriminative Reordering Models for Statistical Machine Translation. In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), Workshop on Statistical Machine Translation*, pages 55–63, New York City, June.

R. Zens and H. Ney. 2006b. N-gram Posterior Probabilities for Statistical Machine Translation. In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), Workshop on Statistical Machine Translation*, pages 72–77, New York City, June.

R. Zens and H. Ney. 2008. Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Honolulu, Hawaii, October.