Joint WMT 2012 Submission of the QUAERO Project

[*]**Markus Freitag**, [*]**Stephan Peitz**, [*]**Matthias Huck**, [*]**Hermann Ney**,
[†]**Jan Niehues**, [†]**Teresa Herrmann**, [†]**Alex Waibel**,
[‡]**Le Hai-son**, [‡]**Thomas Lavergne**, [‡]**Alexandre Allauzen**,
[§]**Bianka Buschbeck**, [§]**Josep Maria Crego**, [§]**Jean Senellart**
[*]RWTH Aachen University, Aachen, Germany
[†]Karlsruhe Institute of Technology, Karlsruhe, Germany
[‡]LIMSI-CNRS, Orsay, France
[§]SYSTRAN Software, Inc.
[*]surname@cs.rwth-aachen.de
[†]firstname.surname@kit.edu
[‡]firstname.lastname@limsi.fr [§]surname@systran.fr

Abstract

This paper describes the joint QUAERO submission to the WMT 2012 machine translation evaluation. Four groups (RWTH Aachen University, Karlsruhe Institute of Technology, LIMSI-CNRS, and SYSTRAN) of the QUAERO project submitted a joint translation for the WMT German→English task. Each group translated the data sets with their own systems and finally the RWTH system combination combined these translations in our final submission. Experimental results show improvements of up to 1.7 points in BLEU and 3.4 points in TER compared to the best single system.

1 Introduction

QUAERO is a European research and development program with the goal of developing multimedia and multilingual indexing and management tools for professional and general public applications (http://www.quaero.org). Research in machine translation is mainly assigned to the four groups participating in this joint submission. The aim of this WMT submission was to show the quality of a joint translation by combining the knowledge of the four project partners. Each group develop and maintain their own different machine translation system. These single systems differ not only in their general approach, but also in the preprocessing of training and test data. To take the advantage of these differences of each translation system, we combined all hypotheses of the different systems, using the RWTH system combination approach.

This paper is structured as follows. In Section 2, the different engines of all four groups are introduced. In Section 3, the RWTH Aachen system combination approach is presented. Experiments with different system selections for system combination are described in Section 4. Finally in Section 5, we discuss the results.

2 Translation Systems

For WMT 2012 each QUAERO partner trained their systems on the parallel Europarl and News Commentary corpora. All single systems were tuned on the newstest2009 or newstest2010 development set. The newstest2011 dev set was used to train the system combination parameters. Finally, the newstest2008-newstest2010 dev sets were used to compare the results of the different system combination settings. In this Section all four different system engines are presented.

2.1 RWTH Aachen Single Systems

For the WMT 2012 evaluation the RWTH utilized RWTH's state-of-the-art phrase-based and hierarchical translation systems. GIZA++ (Och and Ney, 2003) was employed to train word alignments, language models have been created with the SRILM toolkit (Stolcke, 2002).

2.1.1 Phrase-Based System

The phrase-based translation (PBT) system is similar to the one described in Zens and Ney (2008). After phrase pair extraction from the word-aligned parallel corpus, the translation probabilities are estimated by relative frequencies. The standard feature

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 322–329,
Montréal, Canada, June 7-8, 2012.

set also includes an *n*-gram language model, phrase-level IBM-1 and word-, phrase- and distortion-penalties, which are combined in log-linear fashion. The model weights are optimized with standard Mert (Och, 2003) on 200-best lists. The optimization criterium is BLEU.

2.1.2 Hierarchical System

For the hierarchical setups (HPBT) described in this paper, the open source Jane toolkit (Vilar et al., 2010) is 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 model weights are optimized with standard Mert (Och, 2003) on 100-best lists. The optimization criterium is 4BLEU −TER.

2.1.3 Preprocessing

In order to reduce the source vocabulary size translation, the German text was preprocessed by splitting German compound words with the frequency-based method described in (Koehn and Knight, 2003a). To further reduce translation complexity for the phrase-based approach, we performed the long-range part-of-speech based reordering rules proposed by (Popović et al., 2006).

2.1.4 Language Model

For both decoders a 4-gram language model is applied. The language model is trained on the parallel data as well as the provided News crawl, the 10^9 French-English, UN and LDC Gigaword Fourth Edition corpora. For the 10^9 French-English, UN and LDC Gigaword corpora RWTH applied the data selection technique described in (Moore and Lewis, 2010).

2.2 Karlsruhe Institute of Technology Single System

2.2.1 Preprocessing

We preprocess the training data prior to training the system, first by normalizing symbols such as quotes, dashes and apostrophes. Then smart-casing of the first words of each sentence is performed. For the German part of the training corpus we use the hunspell[1] lexicon to learn a mapping from old German spelling to new German spelling to obtain a corpus with homogenous spelling. In addition, we perform compound splitting as described in (Koehn and Knight, 2003b). Finally, we remove very long sentences, empty lines, and sentences that probably are not parallel due to length mismatch.

2.2.2 System Overview

The KIT system uses an in-house phrase-based decoder (Vogel, 2003) to perform translation and optimization with regard to the BLEU score is done using Minimum Error Rate Training as described in Venugopal et al. (2005).

2.2.3 Translation Models

The translation model is trained on the Europarl and News Commentary Corpus and the phrase table is based on a discriminative word alignment (Niehues and Vogel, 2008).

In addition, the system applies a bilingual language model (Niehues et al., 2011) to extend the context of source language words available for translation.

Furthermore, we use a discriminative word lexicon as introduced in (Mauser et al., 2009). The lexicon was trained and integrated into our system as described in (Mediani et al., 2011).

At last, we tried to find translations for out-of-vocabulary (OOV) words by using quasi-morphological operations as described in Niehues and Waibel (2011). For each OOV word, we try to find a related word that we can translate. We modify the ending letters of the OOV word and learn quasi-morphological operations to be performed on the known translation of the related word to synthesize a translation for the OOV word. By this approach we were for example able to translate *Kaminen* into *chimneys* using the known translation *Kamin # chimney*.

2.2.4 Language Models

We use two 4-gram SRI language models, one trained on the News Shuffle corpus and one trained

[1] `http://hunspell.sourceforge.net/`

on the Gigaword corpus. Furthermore, we use a 5-gram cluster-based language model trained on the News Shuffle corpus. The word clusters were created using the MKCLS algorithm. We used 100 word clusters.

2.2.5 Reordering Model

Reordering is performed based on part-of-speech tags obtained using the TreeTagger (Schmid, 1994). Based on these tags we learn probabilistic continuous (Rottmann and Vogel, 2007) and discontinuous (Niehues and Kolss, 2009) rules to cover short and long-range reorderings. The rules are learned from the training corpus and the alignment. In addition, we learned tree-based reordering rules. Therefore, the training corpus was parsed by the Stanford parser (Rafferty and Manning, 2008). The tree-based rules consist of the head node of a subtree and all its children as well as the new order and a probability. These rules were applied recursively. The reordering rules are applied to the source sentences and the reordered sentence variants as well as the original sequence are encoded in a word lattice which is used as input to the decoder. For the test sentences, the reordering based on parts-of-speech and trees allows us to change the word order in the source sentence so that the sentence can be translated more easily. In addition, we build reordering lattices for all training sentences and then extract phrase pairs from the monotone source path as well as from the reordered paths.

2.3 LIMSI-CNRS Single System

LIMSI's system is built with *n*-code (Crego et al., 2011), an open source statistical machine translation system based on bilingual *n*-gram[2]. In this approach, the translation model relies on a specific decomposition of the joint probability of a sentence pair $P(\mathbf{s},\mathbf{t})$ using the *n*-gram assumption: a sentence pair is decomposed into a sequence of bilingual units called *tuples*, defining a joint segmentation of the source and target. In the approach of (Mariño et al., 2006), this segmentation is a by-product of source reordering which ultimately derives from initial word and phrase alignments.

2.3.1 An Overview of *n*-code

The baseline translation model is implemented as a stochastic finite-state transducer trained using a *n*-gram model of (source,target) pairs (Casacuberta and Vidal, 2004). Training this model requires to reorder source sentences so as to match the target word order. This is performed by a stochastic finite-state reordering model, which uses part-of-speech information[3] to generalize reordering patterns beyond lexical regularities.

In addition to the translation model, *eleven* feature functions are combined: a *target-language model*; four *lexicon models*; two *lexicalized reordering models* (Tillmann, 2004) aiming at predicting the orientation of the next translation unit; a 'weak' distance-based *distortion model*; and finally a *word-bonus model* and a *tuple-bonus model* which compensate for the system preference for short translations. The four *lexicon models* are similar to the ones used in a standard phrase based system: two scores correspond to the relative frequencies of the tuples and two lexical weights estimated from the automatically generated word alignments. The weights associated to feature functions are optimally combined using a discriminative training framework (Och, 2003), using the *newstest2009* development set.

The overall search is based on a beam-search strategy on top of a dynamic programming algorithm. Reordering hypotheses are computed in a preprocessing step, making use of reordering rules built from the word reorderings introduced in the tuple extraction process. The resulting reordering hypotheses are passed to the decoder in the form of word lattices (Crego and Mariño, 2007).

2.3.2 Continuous Space Translation Models

One critical issue with standard *n*-gram translation models is that the elementary units are bilingual pairs, which means that the underlying vocabulary can be quite large. Unfortunately, the parallel data available to train these models are typically smaller than the corresponding monolingual corpora used to train target language models. It is very likely then, that such models should face severe estimation problems. In such setting, using neural network language

[2] `http://ncode.limsi.fr/`

[3] Part-of-speech labels for English and German are computed using the TreeTagger (Schmid, 1995).

model techniques seem all the more appropriate. For this study, we follow the recommendations of Le et al. (2012), who propose to factor the joint probability of a sentence pair by decomposing tuples in two (source and target) parts, and further each part in words. This yields a *word factored translation model* that can be estimated in a continuous space using the SOUL architecture (Le et al., 2011).

The design and integration of a SOUL model for large SMT tasks is far from easy, given the computational cost of computing n-gram probabilities. The solution used here was to resort to a two pass approach: the first pass uses a conventional back-off n-gram model to produce a k-best list; in the second pass, the k-best list is reordered using the probabilities of m-gram SOUL translation models. In the following experiments, we used a fixed context size for SOUL of $m = 10$, and used $k = 300$.

2.3.3 Corpora and Data Preprocessing

The parallel data is word-aligned using MGIZA++[4] with default settings. For the English monolingual training data, we used the same setup as last year[5] and thus the same target language model as detailed in (Allauzen et al., 2011).

For English, we took advantage of our in-house text processing tools for tokenization and detokenization steps (Déchelotte et al., 2008) and our system was built in ”true-case”. As German is morphologically more complex than English, the default policy which consists in treating each word form independently is plagued with data sparsity, which is detrimental both at training and decoding time. Thus, the German side was normalized using a specific pre-processing scheme (Allauzen et al., 2010; Durgar El-Kahlout and Yvon, 2010), which notably aims at reducing the lexical redundancy by (i) normalizing the orthography, (ii) neutralizing most inflections and (iii) splitting complex compounds.

2.4 SYSTRAN Software, Inc. Single System

The data submitted by SYSTRAN were obtained by a system composed of the standard SYSTRAN MT engine in combination with a *statistical post editing* (SPE) component.

The SYSTRAN system is traditionally classified as a rule-based system. However, over the decades, its development has always been driven by pragmatic considerations, progressively integrating many of the most efficient MT approaches and techniques. Nowadays, the baseline engine can be considered as a linguistic-oriented system making use of dependency analysis, general transfer rules as well as of large manually encoded dictionaries (100k - 800k entries per language pair).

The SYSTRAN phrase-based SPE component views the output of the rule-based system as the source language, and the (human) reference translation as the target language, see (L. Dugast and Koehn, 2007). It performs corrections and adaptions learned from the 5-gram language model trained on the parallel target-to-target corpus. Moreover, the following measures - limiting unwanted statistical effects - were applied:

- Named entities, time and numeric expressions are replaced by special tokens on both sides. This usually improves word alignment, since the vocabulary size is significantly reduced. In addition, entity translation is handled more reliably by the rule-based engine.
- The intersection of both vocabularies (i.e. vocabularies of the rule-based output and the reference translation) is used to produce an additional parallel corpus to help to improve word alignment.
- Singleton phrase pairs are deleted from the phrase table to avoid overfitting.
- Phrase pairs not containing the same number of entities on the source and the target side are also discarded.

The SPE language model was trained on 2M bilingual phrases from the news/Europarl corpora, provided as training data for *WMT 2012*. An additional language model built from 15M phrases of the English LDC Gigaword corpus using Kneser-Ney (Kneser and Ney, 1995) smoothing was added. Weights for these separate models were tuned by the Mert algorithm provided in the Moses toolkit (P. Koehn et al., 2007), using the provided news development set.

[4] `http://geek.kyloo.net/software`

[5] The fifth edition of the English Gigaword (LDC2011T07) was not used.

3 RWTH Aachen 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. (2006; 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.

4 Experiments

This year, we tried different sets of single systems for system combination. As RWTH has two different translation systems, we put the output of both systems into system combination. Although both systems have the same preprocessing and language model, their hypotheses differ because of their different decoding approach. Compared to the other systems, the system by SYSTRAN has a completely different approach (see section 2.4). It is mainly based on a rule-based system. For the German→English pair, SYSTRAN achieves a lower BLEU score in each test set compared to the other groups. However, since the SYSTRAN system is very different to the others, we still obtain an improvement when we add it also to system combination.

We did experiments with different optimization criteria for the system combination optimization. All results are listed in Table 1 (unoptimized), Table 2 (optimized on BLEU) and Table 3 (optimized on TER-BLEU). Further, we investigated, whether we will loose performance, if a single system is dropped from the system combination. The results show that for each optimization criteria we need all systems to achieve the best results.

For the BLEU optimized system combination, we obtain an improvement compared to the best single systems for all dev sets. For newstest2008, we get an improvement of 1.5 points in BLEU and 1.5 points in TER compared to the best single system of Karlsruhe Institute of Technology. For newstest2009 we get an improvement of 1.9 points in BLEU and 1.5 points in TER compared to the best single system. The system combination of all systems outperforms the best single system with 1.9 points in BLEU and 1.9 points in TER for newstest2010. For newstest2011 the improvement is 1.3 points in BLEU and 2.9 points in TER.

For the TER-BLEU optimized system combination, we achieved more improvement in TER compared to the BLEU optimized system combination. For newstest2008, we get an improvement of 0.8 points in BLEU and 3.0 points in TER compared to the best single system of Karlsruhe Institute of Technology. The system combinations performs better on newstest2009 with 1.3 points in BLEU and 2.7 points in TER. For newstest2010, we get an improvement of 1.7 points in BLEU and 3.4 points in TER and for newstest2011 we get an improvement of 0.7 points in BLEU and 2.5 points in TER.

5 Conclusion

The four statistical machine translation systems of Karlsruhe Institute of Technology, RWTH Aachen and LIMSI and the very structural approach of SYSTRAN produce hypotheses with a huge variability compared to the others. Finally, the RWTH Aachen system combination combined all single system hypotheses to one hypothesis with a higher BLEU and a lower TER score compared to each single system. For each optimization criteria the system combinations using all single systems outperforms the system combinations using one less single system. Although the single system of SYSTRAN has the worst error scores and the RWTH single systems are similar, we achieved the best result in using all single systems. For the WMT 12 evaluation, we submitted the system combination of all systems optimized on BLEU.

Acknowledgments

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

References

Alexandre Allauzen, Josep M. Crego, İlknur Durgar El-Kahlout, and Francois Yvon. 2010. LIMSI's statistical translation systems for WMT'10. In *Proc. of the*

Table 1: All systems for the WMT 2012 German→English translation task (truecase). BLEU and TER results are in percentage. sc denotes system combination. All system combinations are **unoptimized**.

system	**newstest2008**		**newstest2009**		**newstest2010**		**newstest2011**		
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	TER-BLEU
KIT	22.2	61.8	21.3	61.0	24.1	59.0	22.4	60.2	37.9
RWTH.PBT	21.4	62.0	21.3	61.1	23.9	59.1	21.4	61.2	39.7
Limsi	22.2	63.0	22.0	61.8	23.9	59.9	21.8	62.0	40.2
RWTH.HPBT	21.5	62.6	21.5	61.6	23.6	60.2	21.5	61.8	40.4
SYSTRAN	18.3	64.6	17.9	63.4	21.1	60.5	18.3	63.1	44.8
sc-withAllSystems	23.4	59.7	22.9	59.0	26.2	56.5	23.3	58.8	35.5
sc-without-RWTH.PBT	23.2	59.8	22.8	59.0	25.9	56.6	23.1	58.7	35.6
sc-without-RWTH.HPBT	23.2	59.6	22.7	58.9	26.1	56.2	23.1	58.7	35.6
sc-without-Limsi	22.7	60.1	22.4	59.2	25.5	56.7	22.8	58.8	36.0
sc-without-SYSTRAN	23.0	60.3	22.5	59.5	25.7	57.2	23.1	59.2	36.1
sc-without-KIT	23.0	59.9	22.5	59.1	25.9	56.6	22.9	59.1	36.3

Table 2: All systems for the WMT 2012 German→English translation task (truecase). BLEU and TER results are in percentage. sc denotes system combination. All system combinations are **optimized on BLEU** .

system	**newstest2008**		**newstest2009**		**newstest2010**		**newstest2011**		
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	TER-BLEU
sc-withAllSystems	23.7	60.3	23.2	59.5	26.0	57.1	23.7	59.2	35.6
sc-without-RWTH.PBT	23.4	61.1	23.1	59.8	25.5	57.6	23.5	59.5	36.1
sc-without-SYSTRAN	23.3	61.1	22.6	60.5	25.3	58.1	23.5	60.0	36.5
sc-without-Limsi	23.1	60.7	22.6	59.7	25.4	57.5	23.3	59.4	36.2
sc-without-KIT	23.4	60.7	23.0	59.7	25.6	57.7	23.3	59.8	36.5
sc-without-RWTH.HPBT	23.3	59.4	22.8	58.6	26.1	56.0	23.1	58.4	35.2

Table 3: All systems for the WMT 2012 German→English translation task (truecase). BLEU and TER results are in percentage. sc denotes system combination. All system combinations are **optimized on TER-BLEU** .

system	**newstest2008**		**newstest2009**		**newstest2010**		**newstest2011**		
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	TER-BLEU
sc-withAllSystems	23.0	58.8	22.4	58.3	25.8	55.6	23.1	57.7	34.6
sc-without-RWTH.PBT	23.0	59.3	22.5	58.5	25.6	56.0	23.1	58.0	34.9
sc-without-RWTH.HPBT	23.1	59.0	22.6	58.3	25.8	55.6	23.0	58.0	35.0
sc-without-SYSTRAN	22.9	59.7	22.4	59.1	25.6	56.7	23.2	58.5	35.3
sc-without-Limsi	22.7	59.4	22.2	58.7	25.3	56.1	22.7	58.1	35.5
sc-without-KIT	22.9	59.3	22.4	58.5	25.7	55.8	22.7	58.1	35.4

Joint Workshop on Statistical Machine Translation and MetricsMATR, pages 54–59, Uppsala, Sweden.

Alexandre Allauzen, Gilles Adda, Hélène Bonneau-Maynard, Josep M. Crego, Hai-Son Le, Aurélien Max, Adrien Lardilleux, Thomas Lavergne, Artem Sokolov, Guillaume Wisniewski, and François Yvon. 2011. LIMSI @ WMT11. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 309–315, Edinburgh, Scotland, July. Association for Computational Linguistics.

F. Casacuberta and E. Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(3):205–225.

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

J.M. Crego and J.B. Mariño. 2007. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

Josep M. Crego, Franois Yvon, and Jos B. Mario. 2011. N-code: an open-source Bilingual N-gram SMT Toolkit. *Prague Bulletin of Mathematical Linguistics*, 96:49–58.

D. Déchelotte, O. Galibert G. Adda, A. Allauzen, J. Gauvain, H. Meynard, and F. Yvon. 2008. LIMSI's statistical translation systems for WMT'08. In *Proc. of the NAACL-HTL Statistical Machine Translation Workshop*, Columbus, Ohio.

Ilknur Durgar El-Kahlout and Franois Yvon. 2010. The pay-offs of preprocessing for German-English Statistical Machine Translation. In Marcello Federico, Ian Lane, Michael Paul, and Franois Yvon, editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 251–258.

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.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP'95*, pages 181–184, Detroit, MI.

P. Koehn and K. Knight. 2003a. Empirical Methods for Compound Splitting. In *EACL*, Budapest, Hungary.

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

J. Senellart L. Dugast and P. Koehn. 2007. Statistical post-editing on systran's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 220–223, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hai-Son Le, Ilya Oparin, Alexandre Allauzen, Jean-Luc Gauvain, and François Yvon. 2011. Structured output layer neural network language model. In *Proceedings of ICASSP'11*, pages 5524–5527.

Hai-Son Le, Alexandre Allauzen, and François Yvon. 2012. Continuous space translation models with neural networks. In *NAACL '12: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

José B. Mariño, R. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, and M.R. Costa-jussà. 2006. N-gram-based machine translation. *Computational Linguistics*, 32(4).

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. Mari no, 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 *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, Singapore.

Mohammed Mediani, Eunah Cho, Jan Niehues, Teresa Herrmann, and Alex Waibel. 2011. The KIT English-French Translation Systems for IWSLT 2011. In *Proceedings of the Eighth International Workshop on Spoken Language Translation (IWSLT)*.

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. Niehues and M. Kolss. 2009. A POS-Based Model for Long-Range Reorderings in SMT. In *Fourth Workshop on Statistical Machine Translation (WMT 2009)*, Athens, Greece.

J. Niehues and S. Vogel. 2008. Discriminative Word Alignment via Alignment Matrix Modeling. In *Proc. of Third ACL Workshop on Statistical Machine Translation*, Columbus, USA.

Jan Niehues and Alex Waibel. 2011. Using Wikipedia to Translate Domain-specific Terms in SMT. In *Pro-*

ceedings of the Eighth International Workshop on Spoken Language Translation (IWSLT), San Francisco, CA.

Jan Niehues, Teresa Herrmann, Stephan Vogel, and Alex Waibel. 2011. Wider Context by Using Bilingual Language Models in Machine Translation. In *Sixth Workshop on Statistical Machine Translation (WMT 2011)*, Edinburgh, UK.

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.

A. Birch P. Koehn, H. Hoang, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Anna N. Rafferty and Christopher D. Manning. 2008. Parsing three German treebanks: lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*.

K. Rottmann and S. Vogel. 2007. Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model. In *TMI*, Skövde, Sweden.

H. Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In Evelyne Tzoukermann and SusanEditors Armstrong, editors, *Proceedings of the ACL SIGDATWorkshop*, pages 47–50. Kluwer Academic Publishers.

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.

C. Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004*, pages 101–104. Association for Computational Linguistics.

A. Venugopal, A. Zollman, and A. Waibel. 2005. Training and Evaluation Error Minimization Rules for Statistical Machine Translation. In *Workshop on Data-drive Machine Translation and Beyond (WPT-05)*, Ann Arbor, MI.

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.

S. Vogel. 2003. SMT Decoder Dissected: Word Reordering. In *Int. Conf. on Natural Language Processing and Knowledge Engineering*, Beijing, China.

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.