Joint WMT 2013 Submission of the QUAERO Project

[*]Stephan Peitz, [*]Saab Mansour, [*]Matthias Huck, [*]Markus Freitag, [*]Hermann Ney,
[†]Eunah Cho, [†]Teresa Herrmann, [†]Mohammed Mediani, [†]Jan Niehues, [†]Alex Waibel,
[‡]Alexandre Allauzen, [‡]Quoc Khanh Do,
[§]Bianka Buschbeck, [§]Tonio Wandmacher
[*]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 submission of the QUAERO project for the German→English translation task of the *ACL 2013 Eighth Workshop on Statistical Machine Translation* (WMT 2013). The submission was a system combination of the output of four different translation systems provided by RWTH Aachen University, Karlsruhe Institute of Technology (KIT), LIMSI-CNRS and SYSTRAN Software, Inc. The translations were joined using the RWTH's system combination approach. Experimental results show improvements of up to 1.2 points in BLEU and 1.2 points in TER compared to the best single translation.

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 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 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. First, 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. This paper is concluded in Section 5.

2 Translation Systems

For WMT 2013, each QUAERO partner trained their systems on the parallel Europarl (EPPS), News Commentary (NC) corpora and the web-crawled corpus. All single systems were tuned on the newstest2009 *and* newstest2010 development set. The newstest2011 development set was used to tune the system combination parameters. Finally, on newstest2012 the results of the different system combination settings are compared. In this Section, all four different translation engines are presented.

2.1 RWTH Aachen Single System

For the WMT 2013 evaluation, RWTH utilized a phrase-based decoder based on (Wuebker et al., 2012) which is part of RWTH's open-source SMT toolkit Jane 2.1 [1]. GIZA++ (Och and Ney, 2003) was employed to train a word alignment, language models have been created with the SRILM toolkit (Stolcke, 2002).

After phrase pair extraction from the word-aligned parallel corpus, the translation probabilities are estimated by relative frequencies. The standard feature 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. Furthermore, we used an additional reordering model as described in (Galley and Manning, 2008). By this model six

[1] http://www-i6.informatik.rwth-aachen.de/jane/

additional feature are added to the log-linear combination. The model weights are optimized with standard Mert (Och, 2003a) on 200-best lists. The optimization criterion is BLEU.

2.1.1 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, 2003). 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).

2.1.2 Translation Model

We applied filtering and weighting for domain-adaptation similarly to (Mansour et al., 2011) and (Mansour and Ney, 2012). For filtering the bilingual data, a combination of LM and IBM Model 1 scores was used. In addition, we performed weighted phrase extraction by using a combined LM and IBM Model 1 weight.

2.1.3 Language Model

During decoding 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.

2.2 Karlsruhe Institute of Technology Single System

2.2.1 Preprocessing

The training data was preprocessed prior to the training. Symbols such as quotes, dashes and apostrophes are normalized. Then the first words of each sentence are smart-cased. For the German part of the training corpus, the hunspell[2] lexicon was used, in order to learn a mapping from old German spelling to new German writing rules. Compound-splitting was also performed as described in Koehn and Knight (2003). We also removed very long sentences, empty lines, and sentences which show big mismatch on the length.

2.2.2 Filtering

The web-crawled corpus was filtered using an SVM classifier as described in (Mediani et al., 2011). The lexica used in this filtering task were obtained from Giza alignments trained on the cleaner corpora, EPPS and NC. Assuming that this corpus is very noisy, we biased our classifier more towards precision than recall. This was realized by giving higher number of false examples (80% of the training data).

This filtering technique ruled out more than 38% of the corpus (the unfiltered corpus contains around 2.4M pairs, 0.9M of which were rejected in the filtering task).

2.2.3 System Overview

The in-house phrase-based decoder (Vogel, 2003) is used to perform decoding. Optimization with regard to the BLEU score is done using Minimum Error Rate Training (MERT) as described in Venugopal et al. (2005).

2.2.4 Reordering Model

We applied part-of-speech (POS) based reordering using probabilistic continuous (Rottmann and Vogel, 2007) and discontinuous (Niehues and Kolss, 2009) rules. This was learned using POS tags generated by the TreeTagger (Schmid, 1994) for short and long range reorderings respectively.

In addition to this POS-based reordering, we also used tree-based reordering rules. Syntactic parse trees of the whole training corpus and the word alignment between source and target language are used to learn rules on how to reorder the constituents in a German source sentence to make it match the English target sentence word order better (Herrmann et al., 2013). The training corpus was parsed by the Stanford parser (Rafferty and Manning, 2008). 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.

Moreover, our reordering model was extended so that it could include the features of lexicalized reordering model. The reordering probabilities for each phrase pair are stored as well as the original position of each word in the lattice. During the decoding, the reordering origin of the words is checked along with its probability added as an additional score.

2.2.5 Translation Models

The translation model uses the parallel data of EPPS, NC, and the filtered web-crawled data. As word alignment, we used the Discriminative Word Alignment (DWA) as shown in (Niehues and Vo-

[2] http://hunspell.sourceforge.net/

gel, 2008). The phrase pairs were extracted using different source word order suggested by the POS-based reordering models presented previously as described in (Niehues et al., 2009).

In order to extend the context of source language words, we applied a bilingual language model (Niehues et al., 2011). A Discriminative Word Lexicon (DWL) introduced in (Mauser et al., 2009) was extended so that it could take the source context also into the account. For this, we used a bag-of-ngrams instead of representing the source sentence as a bag-of-words. Filtering based on counts was then applied to the features for higher order n-grams. In addition to this, the training examples were created differently so that we only used the words that occur in the n-best list but not in the reference as negative example.

2.2.6 Language Models

We build separate language models and combined them prior to decoding. As word-token based language models, one language model is built on EPPS, NC, and giga corpus, while another one is built using crawled data. We combined the LMs linearly by minimizing the perplexity on the development data. As a bilingual language model we used the EPPS, NC, and the web-crawled data and combined them. Furthermore, we use a 5-gram cluster-based language model with 1,000 word clusters, which was trained on the EPPS and NC corpus. The word clusters were created using the MKCLS algorithm.

2.3 LIMSI-CNRS Single System

2.3.1 System overview

LIMSI's system is built with *n*-code (Crego et al., 2011), an open source statistical machine translation system based on bilingual *n*-gram[3]. In this approach, the translation model relies on a specific decomposition of the joint probability of a sentence pair 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.

2.3.2 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[4] 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 use 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, 2003b).

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 Mario, 2006).

2.3.3 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, even for small translation tasks. Unfortunately, the parallel data available to train these models are typically order of magnitudes 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 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

[3] http://ncode.limsi.fr/

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

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 $k = 300$.

2.3.4 Corpora and data pre-processing

All the parallel data allowed in the constrained task are pooled together to create a single parallel corpus. This corpus is word-aligned using MGIZA++[5] with default settings. For the English monolingual training data, we used the same setup as last year[6] and thus the same target language model as detailed in (Allauzen et al., 2011).

For English, we also took advantage of our in-house text processing tools for the tokenization and detokenization steps (Dchelotte et al., 2008) and our system is 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 (described in (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

In the past few years, SYSTRAN has been focusing on the introduction of statistical approaches to its rule-based backbone, leading to *Hybrid Machine Translation*.

The technique of *Statistical Post-Editing* (Dugast et al., 2007) is used to automatically edit the output of the rule-based system. A Statistical Post-Editing (SPE) module is generated from a bilingual corpus. It is basically a translation module by itself, however it is trained on rule-based translations and reference data. It applies corrections and adaptations learned from a phrase-based 5-gram language model. Using this two-step process will implicitly keep long distance relations and other constraints determined by the rule-based system while significantly improving phrasal fluency. It has the advantage that quality improvements can be achieved with very little but targeted bilingual data, thus significantly reducing training time and increasing translation performance.

The basic setup of the SPE component is identical to the one described in (Dugast et al., 2007). A statistical translation model is trained on the rule-based translation of the source and the target side of the parallel corpus. Language models are trained on each target half of the parallel corpora and also on additional in-domain corpora. Moreover, the following measures - limiting unwanted statistical effects - were applied:

- Named entities 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 (whose target is identical to the source). This was added to the parallel text in order 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.
- Phrase pairs appearing less than 2 times were pruned.

The SPE language model was trained on 2M phrases from the news/europarl and CommonCrawl corpora, provided as training data for *WMT 2013*. Weights for these separate models were tuned by the Mert algorithm provided in the Moses toolkit (Koehn et al., 2007), using the provided news development set.

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

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

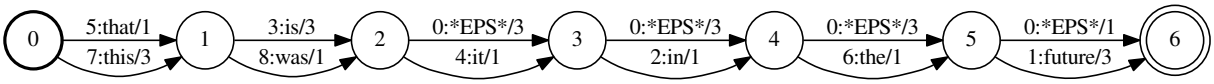

Figure 1: Confusion network of four different hypotheses.

3 RWTH Aachen 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, each of the given single systems generates one confusion network with its own 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 final network for one source sentence is the union of all confusion networks generated from the different primary systems. That allows the system combination to select the word order from different system outputs.

Before performing system combination, each translation output was normalized by tokenization and lowercasing. The output of the combination was then truecased based on the original truecased output.

The model weights of the system combination are optimized with standard Mert (Och, 2003a) on 100-best lists. We add one voting feature for each single system to the log-linear framework of the system combination. The voting feature fires for each word the single system agrees on. Moreover, 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 are combined in this log-linear model. The optimization criterion is 4BLEU-TER.

4 Experimental Results

In this year's experiments, we tried to improve the result of the system combination further by combining single systems tuned on different development sets. The idea is to achieve a more stable performance in terms of translation quality, if the single systems are not optimized on the same data set. In Table 1, the results of each provided single system tuned on newstest2009 and newstest2010 are shown. For RWTH, LIMSI and SYSTRAN, it seems that the performance of the single system depends on the chosen tuning set. However, the translation quality of the single systems provided by KIT is stable.

Table 1: Comparison of single systems tuned on newstest2009 and newstest2010. The results are reported on newstest2012.

single systems tuned on		newstest2012	
	newstest	BLEU	TER
KIT	2009	24.6	58.4
	2010	24.6	58.6
LIMSI	2009	22.5	61.5
	2010	22.6	59.8
SYSTRAN	2009	20.9	63.3
	2010	21.2	62.2
RWTH	2009	23.7	60.8
	2010	24.4	58.8

As initial approach and for the final submission, we grouped single systems with dissimilar approaches. Thus, KIT (phrase-based SMT) and SYSTRAN (rule-based MT) tuned their system on newstest2010, while RWTH (phrase-based SMT) and LIMSI (n-gram) optimized on newstest2009.

To compare the impact of this approach, all possible combinations were checked (Table 2). However, it seems that the translation quality can not be improved by this approach. For the test set (newstest2012), BLEU is steady around 25.6 points. Even if the single system with lowest BLEU are combined (KIT 2010, LIMSI 2009, SYSTRAN 2010, RWTH 2009), the translation quality in terms of BLEU is comparable with the combination of the best single systems (KIT 2009, LIMSI 2010, SYSTRAN 2010, RWTH 2010). However, we could gain 1.0 point in TER.

Due to the fact, that for the final submission the initial grouping was available only, we kept this

Table 2: Comparison of different system combination settings. For each possible combination of systems tuned on different tuning sets, a system combination was set up, re-tuned on newstest2011 and evaluated on newstest2012. The setting used for further experiments is set in boldface.

single systems				system combinations			
KIT	LIMSI	SYSTRAN	RWTH	newstest2011		newstest2012	
tuned on newstest				BLEU	TER	BLEU	TER
2009	2009	2009	2009	24.6	58.0	25.6	56.8
2010	2010	2010	2010	24.2	58.1	25.6	57.7
2010	2009	2009	2009	24.5	57.9	25.7	57.4
2009	2010	2009	2009	24.4	58.3	25.7	57.0
2009	2009	2010	2009	24.5	57.9	25.6	57.0
2009	2009	2009	2010	24.5	58.0	25.6	56.8
2009	2010	2010	2010	24.1	57.5	25.4	56.4
2010	2009	2010	2010	24.3	57.6	25.6	56.9
2010	2010	2009	2010	24.2	58.0	25.6	57.3
2010	2010	2010	2009	24.3	57.9	25.5	57.6
2010	2010	2009	2009	24.4	58.1	25.6	57.5
2009	2009	2010	2010	24.4	57.8	25.5	56.6
2009	2010	2010	2009	24.4	58.2	25.5	57.0
2009	2010	2009	2010	24.2	57.8	25.5	56.8
2010	2009	2009	2010	24.4	57.9	25.6	57.4
2010	**2009**	**2010**	**2009**	**24.4**	**57.7**	**25.6**	**57.4**

Table 3: Results of the final submission (boldface) compared with best single system on newstest2012.

	newstest2011		newstest2012	
	BLEU	TER	BLEU	TER
best single	23.2	60.9	24.6	58.4
system comb.	24.4	57.7	25.6	57.4
+ IBM-1	24.6	58.1	25.6	57.6
+ bigLM	**24.6**	**57.9**	**25.8**	**57.2**

combination. To improve this baseline further, two additional models were added. We applied lexical smoothing (*IBM-1*) and an additional language model (*bigLM*) trained on the English side of the parallel data and the News shuffle corpus. The results are presented in Table 3.

The baseline was slightly improved by 0.2 points in BLEU and TER. Note, this system combination was the final submission.

5 Conclusion

For the participation in the WMT 2013 shared translation task, the partners of the QUAERO project (Karlsruhe Institute of Technology, RWTH Aachen University, LIMSI-CNRS and SYSTRAN Software, Inc.) provided a joint submission. By joining the output of four different translation systems with RWTH's system combination, we reported an improvement of up to 1.2 points in BLEU and TER.

Combining systems optimized on different tuning sets does not seem to improve the translation quality. However, by adding additional model, the baseline was slightly improved.

All in all, we conclude that the variability in terms of BLEU does not influence the final result. It seems that using different approaches of MT in a system combination is more important (Freitag et al., 2012).

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 François Yvon. 2010. LIMSI's statistical translation systems for WMT'10. In *Proc. of*

the 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.

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

Josep M. Crego and José B. Mario. 2006. 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.

Loïc Dugast, Jean Senellart, and Philipp 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.

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.

Daniel Dchelotte, Gilles Adda, Alexandre Allauzen, Olivier Galibert, Jean-Luc Gauvain, Hlne Maynard, and Franois Yvon. 2008. LIMSI's statistical translation systems for WMT'08. In *Proc. of the NAACL-HTL Statistical Machine Translation Workshop*, Columbus, Ohio.

Markus Freitag, Stephan Peitz, Matthias Huck, Hermann Ney, Teresa Herrmann, Jan Niehues, Alex Waibel, Alexandre Allauzen, Gilles Adda, Bianka Buschbeck, Josep Maria Crego, and Jean Senellart. 2012. Joint wmt 2012 submission of the quaero project. In *NAACL 2012 Seventh Workshop on Statistical Machine Translation*, pages 322–329, Montreal, Canada, June.

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, October. Association for Computational Linguistics.

Teresa Herrmann, Jan Niehues, and Alex Waibel. 2013. Combining Word Reordering Methods on different Linguistic Abstraction Levels for Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, Altanta, Georgia, USA, June. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *EACL*, Budapest, Hungary.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantine, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. pages 177–180, Prague, Czech Republic, June.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. pages 228–231, Prague, Czech Republic, June.

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*.

Saab Mansour and Hermann Ney. 2012. A simple and effective weighted phrase extraction for machine translation adaptation. In *International Workshop on Spoken Language Translation*, pages 193–200, Hong Kong, December.

Sab 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)*, San Francisco, CA, December.

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrick Lambert, José A.R. Fonollosa, and Marta R. Costa-Jussà. 2006. N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549.

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)*.

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

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

Jan Niehues, Teresa Herrmann, Muntsin Kolss, and Alex Waibel. 2009. The Universität Karlsruhe Translation System for the EACL-WMT 2009. In *Fourth Workshop on Statistical Machine Translation (WMT 2009)*, Athens, Greece.

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.

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. 2003a. 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.

Franz Josef Och. 2003b. Minimum error rate training in statistical machine translation. In *ACL '03: Proc. of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167.

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*.

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

Helmut 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 SIGDAT-Workshop*, pages 47–50. Kluwer Academic Publishers.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. Int. Conf. on Spoken Language Processing*, volume 2, pages 901–904, Denver, Colorado, USA.

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

Ashish Venugopal, Andreas Zollman, and Alex 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.

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

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.