EU-BRIDGE MT: Combined Machine Translation

*Markus Freitag, *Stephan Peitz, *Joern Wuebker, *Hermann Ney,
‡Matthias Huck, ‡Rico Sennrich, ‡Nadir Durrani,
‡Maria Nadejde, ‡Philip Williams, ‡Philipp Koehn,
†Teresa Herrmann, †Eunah Cho, †Alex Waibel
*RWTH Aachen University, Aachen, Germany
‡University of Edinburgh, Edinburgh, Scotland
†Karlsruhe Institute of Technology, Karlsruhe, Germany
*{freitag,peitz,wuebker,ney}@cs.rwth-aachen.de
‡{mhuck,ndurrani,pkoehn}@inf.ed.ac.uk
‡v1rsennr@staffmail.ed.ac.uk
‡maria.nadejde@gmail.com,p.j.williams-2@sms.ed.ac.uk
†{teresa.herrmann,eunah.cho,alex.waibel}@kit.edu

Abstract

This paper describes one of the collaborative efforts within EU-BRIDGE to further advance the state of the art in machine translation between two European language pairs, German→English and English→German. Three research institutes involved in the EU-BRIDGE project combined their individual machine translation systems and participated with a joint setup in the shared translation task of the evaluation campaign at the *ACL 2014 Eighth Workshop on Statistical Machine Translation* (WMT 2014).

We combined up to nine different machine translation engines via system combination. RWTH Aachen University, the University of Edinburgh, and Karlsruhe Institute of Technology developed several individual systems which serve as system combination input. We devoted special attention to building syntax-based systems and combining them with the phrase-based ones. The joint setups yield empirical gains of up to 1.6 points in BLEU and 1.0 points in TER on the WMT newstest2013 test set compared to the best single systems.

1 Introduction

EU-BRIDGE[1] is a European research project which is aimed at developing innovative speech translation technology. This paper describes a joint WMT submission of three EU-BRIDGE project partners. RWTH Aachen University (RWTH), the University of Edinburgh (UEDIN) and Karlsruhe Institute of Technology (KIT) all provided several individual systems which were combined by means of the RWTH Aachen system combination approach (Freitag et al., 2014). As distinguished from our EU-BRIDGE joint submission to the IWSLT 2013 evaluation campaign (Freitag et al., 2013), we particularly focused on translation of news text (instead of talks) for WMT. Besides, we put an emphasis on engineering syntax-based systems in order to combine them with our more established phrase-based engines. We built combined system setups for translation from German to English as well as from English to German. This paper gives some insight into the technology behind the system combination framework and the combined engines which have been used to produce the joint EU-BRIDGE submission to the WMT 2014 translation task.

The remainder of the paper is structured as follows: We first describe the individual systems by RWTH Aachen University (Section 2), the University of Edinburgh (Section 3), and Karlsruhe Institute of Technology (Section 4). We then present the techniques for machine translation system combination in Section 5. Experimental results are given in Section 6. We finally conclude the paper with Section 7.

2 RWTH Aachen University

RWTH (Peitz et al., 2014) employs both the phrase-based (*RWTH scss*) and the hierarchical (*RWTH hiero*) decoder implemented in RWTH's publicly available translation toolkit Jane (Vilar

[1] http://www.eu-bridge.eu

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 105–113,
Baltimore, Maryland USA, June 26–27, 2014. ©2014 Association for Computational Linguistics

et al., 2010; Wuebker et al., 2012). The model weights of all systems have been tuned with standard Minimum Error Rate Training (Och, 2003) on a concatenation of the newstest2011 and newstest2012 sets. RWTH used BLEU as optimization objective. Both for language model estimation and querying at decoding, the KenLM toolkit (Heafield et al., 2013) is used. All RWTH systems include the standard set of models provided by Jane. Both systems have been augmented with a hierarchical orientation model (Galley and Manning, 2008; Huck et al., 2013) and a cluster language model (Wuebker et al., 2013). The phrase-based system (*RWTH scss*) has been further improved by maximum expected BLEU training similar to (He and Deng, 2012). The latter has been performed on a selection from the News Commentary, Europarl and Common Crawl corpora based on language and translation model cross-entropies (Mansour et al., 2011).

3 University of Edinburgh

UEDIN contributed phrase-based and syntax-based systems to both the German→English and the English→German joint submission.

3.1 Phrase-based Systems

UEDIN's phrase-based systems (Durrani et al., 2014) have been trained using the Moses toolkit (Koehn et al., 2007), replicating the settings described in (Durrani et al., 2013b). The features include: a maximum sentence length of 80, grow-diag-final-and symmetrization of GIZA++ alignments, an interpolated Kneser-Ney smoothed 5-gram language model with KenLM (Heafield, 2011) used at runtime, a lexically-driven 5-gram operation sequence model (OSM) (Durrani et al., 2013a), msd-bidirectional-fe lexicalized reordering, sparse lexical and domain features (Hasler et al., 2012), a distortion limit of 6, a maximum phrase length of 5, 100-best translation options, Minimum Bayes Risk decoding (Kumar and Byrne, 2004), cube pruning (Huang and Chiang, 2007), with a stack size of 1000 during tuning and 5000 during testing and the no-reordering-over-punctuation heuristic. UEDIN uses POS and morphological target sequence models built on the in-domain subset of the parallel corpus using Kneser-Ney smoothed 7-gram models as additional factors in phrase translation models (Koehn and Hoang, 2007). UEDIN has furthermore built OSM models over POS and morph sequences following Durrani et al. (2013c). The English→German system additionally comprises a target-side LM over automatically built word classes (Birch et al., 2013). UEDIN has applied syntactic pre-reordering (Collins et al., 2005) and compound splitting (Koehn and Knight, 2003) of the source side for the German→English system. The systems have been tuned on a very large tuning set consisting of the test sets from 2008-2012, with a total of 13,071 sentences. UEDIN used newstest2013 as held-out test set. On top of *UEDIN phrase-based 1* system, *UEDIN phrase-based 2* augments word classes as additional factor and learns an interpolated target sequence model over cluster IDs. Furthermore, it learns OSM models over POS, morph and word classes.

3.2 Syntax-based Systems

UEDIN's syntax-based systems (Williams et al., 2014) follow the GHKM syntax approach as proposed by **G**alley, **H**opkins, **K**night, and **M**arcu (Galley et al., 2004). The open source *Moses* implementation has been employed to extract GHKM rules (Williams and Koehn, 2012). Composed rules (Galley et al., 2006) are extracted in addition to minimal rules, but only up to the following limits: at most twenty tree nodes per rule, a maximum depth of five, and a maximum size of five. Singleton hierarchical rules are dropped.

The features for the syntax-based systems comprise Good-Turing-smoothed phrase translation probabilities, lexical translation probabilities in both directions, word and phrase penalty, a rule rareness penalty, a monolingual PCFG probability, and a 5-gram language model. UEDIN has used the SRILM toolkit (Stolcke, 2002) to train the language model and relies on KenLM for language model scoring during decoding. Model weights are optimized to maximize BLEU. 2000 sentences from the newstest2008-2012 sets have been selected as a development set. The selected sentences obtained high sentence-level BLEU scores when being translated with a baseline phrase-based system, and each contain less than 30 words for more rapid tuning. Decoding for the syntax-based systems is carried out with cube pruning using Moses' hierarchical decoder (Hoang et al., 2009).

UEDIN's German→English syntax-based setup is a string-to-tree system with compound splitting

on the German source-language side and syntactic annotation from the Berkeley Parser (Petrov et al., 2006) on the English target-language side.

For English→German, UEDIN has trained various string-to-tree GHKM syntax systems which differ with respect to the syntactic annotation. A tree-to-string system and a string-to-string system (with rules that are not syntactically decorated) have been trained as well. The English→German UEDIN GHKM system names in Table 3 denote:

UEDIN GHKM S2T (ParZu): A string-to-tree system trained with target-side syntactic annotation obtained with ParZu (Sennrich et al., 2013). It uses a modified syntactic label set, target-side compound splitting, and additional syntactic constraints.

UEDIN GHKM S2T (BitPar): A string-to-tree system trained with target-side syntactic annotation obtained with BitPar (Schmid, 2004).

UEDIN GHKM S2T (Stanford): A string-to-tree system trained with target-side syntactic annotation obtained with the German Stanford Parser (Rafferty and Manning, 2008a).

UEDIN GHKM S2T (Berkeley): A string-to-tree system trained with target-side syntactic annotation obtained with the German Berkeley Parser (Petrov and Klein, 2007; Petrov and Klein, 2008).

UEDIN GHKM T2S (Berkeley): A tree-to-string system trained with source-side syntactic annotation obtained with the English Berkeley Parser (Petrov et al., 2006).

UEDIN GHKM S2S (Berkeley): A string-to-string system. The extraction is GHKM-based with syntactic target-side annotation from the German Berkeley Parser, but we strip off the syntactic labels. The final grammar contains rules with a single generic non-terminal instead of syntactic ones, plus rules that have been added from plain phrase-based extraction (Huck et al., 2014).

4 Karlsruhe Institute of Technology

The KIT translations (Herrmann et al., 2014) are generated by an in-house phrase-based translations system (Vogel, 2003). The provided News Commentary, Europarl, and Common Crawl parallel corpora are used for training the translation model. The monolingual part of those parallel corpora, the News Shuffle corpus for both directions and additionally the Gigaword corpus for German→English are used as monolingual training data for the different language models. Optimization is done with Minimum Error Rate Training as described in (Venugopal et al., 2005), using newstest2012 and newstest2013 as development and test data respectively.

Compound splitting (Koehn and Knight, 2003) is performed on the source side of the corpus for German→English translation before training. In order to improve the quality of the web-crawled Common Crawl corpus, noisy sentence pairs are filtered out using an SVM classifier as described by Mediani et al. (2011).

The word alignment for German→English is generated using the GIZA++ toolkit (Och and Ney, 2003). For English→German, KIT uses discriminative word alignment (Niehues and Vogel, 2008). Phrase extraction and scoring is done using the Moses toolkit (Koehn et al., 2007). Phrase pair probabilities are computed using modified Kneser-Ney smoothing as in (Foster et al., 2006).

In both systems KIT applies short-range reorderings (Rottmann and Vogel, 2007) and long-range reorderings (Niehues and Kolss, 2009) based on POS tags (Schmid, 1994) to perform source sentence reordering according to the target language word order. The long-range reordering rules are applied to the training corpus to create reordering lattices to extract the phrases for the translation model. In addition, a tree-based reordering model (Herrmann et al., 2013) trained on syntactic parse trees (Rafferty and Manning, 2008b; Klein and Manning, 2003) as well as a lexicalized reordering model (Koehn et al., 2005) are applied.

Language models are trained with the SRILM toolkit (Stolcke, 2002) and use modified Kneser-Ney smoothing. Both systems utilize a language model based on automatically learned word classes using the MKCLS algorithm (Och, 1999). The English→German system comprises language models based on fine-grained part-of-speech tags (Schmid and Laws, 2008). In addition, a bilingual language model (Niehues et al., 2011) is used as well as a discriminative word lexicon (Mauser et al., 2009) using source context to guide the word choices in the target sentence.

In total, the English→German system uses the following language models: two 4-gram word-based language models trained on the parallel data and the filtered Common Crawl data separately, two 5-gram POS-based language models trained on the same data as the word-based language models, and a 4-gram cluster-based language model trained on 1,000 MKCLS word classes.

The German→English system uses a 4-gram word-based language model trained on all monolingual data and an additional language model trained on automatically selected data (Moore and Lewis, 2010). Again, a 4-gram cluster-based language model trained on 1000 MKCLS word classes is applied.

5 System Combination

System combination is used to produce consensus translations from multiple hypotheses which are outputs of different translation engines. The consensus translations can be better in terms of translation quality than any of the individual hypotheses. To combine the engines of the project partners for the EU-BRIDGE joint setups, we apply a system combination implementation that has been developed at RWTH Aachen University.

The implementation of RWTH's approach to machine translation system combination is described in (Freitag et al., 2014). This approach includes an enhanced alignment and reordering framework. Alignments between the system outputs are learned using METEOR (Banerjee and Lavie, 2005). A confusion network is then built using one of the hypotheses as "primary" hypothesis. We do not make a hard decision on which of the hypotheses to use for that, but instead combine all possible confusion networks into a single lattice. Majority voting on the generated lattice is performed using the prior probabilities for each system as well as other statistical models, e.g. a special n-gram language model which is learned on the input hypotheses. Scaling factors of the models are optimized using the Minimum Error Rate Training algorithm. The translation with the best total score within the lattice is selected as consensus translation.

6 Results

In this section, we present our experimental results on the two translation tasks, German→English and English→German. The weights of the individual system engines have been optimized on different test sets which partially or fully include newstest2011 or newstest2012. System combination weights are either optimized on newstest2011 or newstest2012. We kept newstest2013 as an unseen test set which has not been used for tuning the system combination or any of the individual systems.

6.1 German→English

The automatic scores of all individual systems as well as of our final system combination submission are given in Table 1. KIT, UEDIN and RWTH are each providing one individual phrase-based system output. RWTH (*hiero*) and UEDIN (*GHKM*) are providing additional systems based on the hierarchical translation model and a string-to-tree syntax model. The pairwise difference of the single system performances is up to 1.3 points in BLEU and 2.5 points in TER. For German→English, our system combination parameters are optimized on newstest2012. System combination gives us a gain of 1.6 points in BLEU and 1.0 points in TER for newstest2013 compared to the best single system.

In Table 2 the pairwise BLEU scores for all individual systems as well as for the system combination output are given. The pairwise BLEU score of both RWTH systems (taking one as hypothesis and the other one as reference) is the highest for all pairs of individual system outputs. A high BLEU score means similar hypotheses. The syntax-based system of UEDIN and *RWTH scss* differ mostly, which can be observed from the fact of the lowest pairwise BLEU score. Furthermore, we can see that better performing individual systems have higher BLEU scores when evaluating against the system combination output.

In Figure 1 system combination output is compared to the best single system *KIT*. We distribute the sentence-level BLEU scores of all sentences of newstest2013. To allow for sentence-wise evaluation, all bi-, tri-, and four-gram counts are initialized with 1 instead of 0. Many sentences have been improved by system combination. Nevertheless, some sentences fall off in quality compared to the individual system output of *KIT*.

6.2 English→German

The results of all English→German system setups are given in Table 3. For the English→German translation task, only UEDIN and KIT are con-

system	newstest2011		newstest2012		newstest2013	
	BLEU	TER	BLEU	TER	BLEU	TER
KIT	25.0	57.6	25.2	57.4	27.5	54.4
UEDIN	23.9	59.2	24.7	58.3	27.4	55.0
RWTH scss	23.6	59.5	24.2	58.5	27.0	55.0
RWTH hiero	23.3	59.9	24.1	59.0	26.7	55.9
UEDIN GHKM S2T (Berkeley)	23.0	60.1	23.2	60.8	26.2	56.9
syscom	25.6	57.1	**26.4**	**56.5**	**29.1**	**53.4**

Table 1: Results for the German→English translation task. The system combination is tuned on newstest2012, newstest2013 is used as held-out test set for all individual systems and system combination. Bold font indicates system combination results that are significantly better than the best single system with $p < 0.05$.

	KIT	UEDIN	RWTH scss	RWTH hiero	UEDIN S2T	syscom
KIT		59.07	57.60	57.91	55.62	77.68
UEDIN	59.17		56.96	57.84	59.89	72.89
RWTH scss	57.64	56.90		64.94	53.10	71.16
RWTH hiero	57.98	57.80	64.97		55.73	70.87
UEDIN S2T	55.75	59.95	53.19	55.82		65.35
syscom	77.76	72.83	71.17	70.85	65.24	

Table 2: Cross BLEU scores for the German→English newstest2013 test set. (Pairwise BLEU scores: each entry is taking the horizontal system as hypothesis and the other one as reference.)

system	newstest2011		newstest2012		newstest2013	
	BLEU	TER	BLEU	TER	BLEU	TER
UEDIN phrase-based 1	17.5	67.3	18.2	65.0	20.5	62.7
UEDIN phrase-based 2	17.8	66.9	18.5	64.6	20.8	62.3
UEDIN GHKM S2T (ParZu)	17.2	67.6	18.0	65.5	20.2	62.8
UEDIN GHKM S2T (BitPar)	16.3	69.0	17.3	66.6	19.5	63.9
UEDIN GHKM S2T (Stanford)	16.1	69.2	17.2	67.0	19.0	64.2
UEDIN GHKM S2T (Berkeley)	16.3	68.9	17.2	66.7	19.3	63.8
UEDIN GHKM T2S (Berkeley)	16.7	68.9	17.5	66.9	19.5	63.8
UEDIN GHKM S2S (Berkeley)	16.3	69.2	17.3	66.8	19.1	64.3
KIT	17.1	67.0	17.8	64.8	20.2	62.2
syscom	**18.4**	**65.0**	18.7	**63.4**	*21.3*	**60.6**

Table 3: Results for the English→German translation task. The system combination is tuned on newstest2011, newstest2013 is used as held-out test set for all individual systems and system combination. Bold font indicates system combination results that are significantly (Bisani and Ney, 2004) better than the best single system with $p < 0.05$. Italic font indicates system combination results that are significantly better than the best single system with $p < 0.1$.

tributing individual systems. KIT is providing a phrase-based system output, UEDIN is providing two phrase-based system outputs and six syntax-based ones (*GHKM*). For English→German, our system combination parameters are optimized on newstest2011. Combining all nine different system outputs yields an improvement of 0.5 points in BLEU and 1.7 points in TER over the best single system performance.

In Table 4 the cross BLEU scores for all English→German systems are given. The individual system of *KIT* and the syntax-based *ParZu* system of UEDIN have the lowest BLEU score when scored against each other. Both approaches are quite different and both are coming from different institutes. In contrast, both phrase-based systems *pbt 1* and *pbt 2* from UEDIN are very similar and hence have a high pairwise BLEU score.

	pbt 1	pbt 2	ParZu	BitPar	Stanford	S2T	T2S	S2S	KIT	syscom
pbt 1		75.84	51.61	53.93	55.32	54.79	54.52	60.92	54.80	70.12
pbt 2	75.84		51.96	53.39	53.93	53.97	53.10	57.32	54.04	73.75
ParZu	51.57	51.91		56.67	55.11	56.05	52.13	51.22	48.14	68.39
BitPar	54.00	53.45	56.78		64.59	65.67	56.33	56.62	49.23	62.08
Stanford	55.37	53.98	55.19	64.56		69.22	58.81	61.19	50.50	61.51
S2T	54.83	54.02	56.14	65.64	69.21		59.32	60.16	50.07	62.81
T2S	54.57	53.15	52.21	56.30	58.81	59.32		59.34	50.01	63.13
S2S	60.96	57.36	51.29	56.59	61.18	60.15	59.33		53.68	60.46
KIT	54.75	53.98	48.13	49.13	50.41	49.98	49.93	53.59		63.33
syscom	70.01	73.63	68.32	61.92	61.37	62.67	62.99	60.32	63.27	

Table 4: Cross BLEU scores for the German→English newstest2013 test set. (Pairwise BLEU scores: each entry is taking the horizontal system as reference and the other one as hypothesis.)

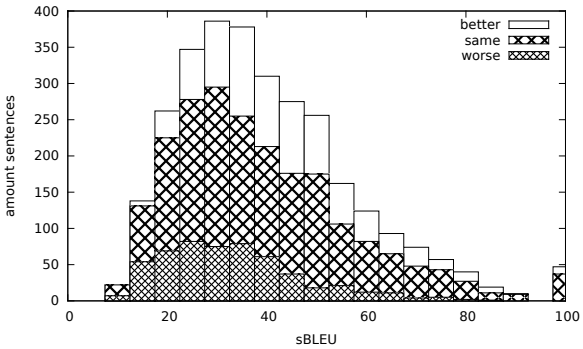

Figure 1: Sentence distribution for the German→English newstest2013 test set comparing system combination output against the best individual system.

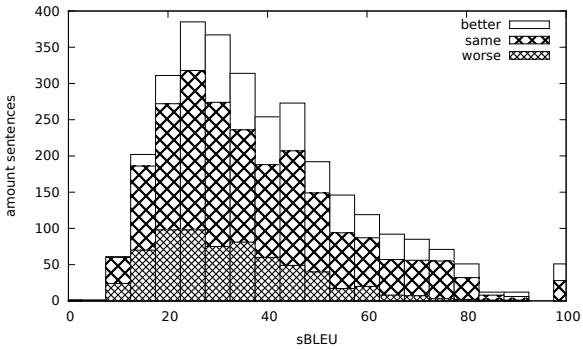

Figure 2: Sentence distribution for the English→German newstest2013 test set comparing system combination output against the best individual system.

As for the German→English translation direction, the best performing individual system outputs are also having the highest BLEU scores when evaluated against the final system combination output.

In Figure 2 system combination output is compared to the best single system *pbt 2*. We distribute the sentence-level BLEU scores of all sentences of newstest2013. Many sentences have been improved by system combination. But there is still room for improvement as some sentences are still better in terms of sentence-level BLEU in the individual best system *pbt 2*.

7 Conclusion

We achieved significantly better translation performance with gains of up to +1.6 points in BLEU and -1.0 points in TER by combining up to nine different machine translation systems. Three different research institutes (RWTH Aachen University, University of Edinburgh, Karlsruhe Institute of Technology) provided machine translation engines based on different approaches like phrase-based, hierarchical phrase-based, and syntax-based. For English→German, we included six different syntax-based systems, which were combined to our final combined translation. The automatic scores of all submitted system outputs for the actual 2014 evaluation set are presented on the WMT submission page.[2] Our joint submission is the best submission in terms of BLEU and TER for both translation directions German→English and English→German without adding any new data.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287658.

Rico Sennrich has received funding from the Swiss National Science Foundation under grant P2ZHP1_148717.

[2] http://matrix.statmt.org/

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI, USA, June.

Alexandra Birch, Nadir Durrani, and Philipp Koehn. 2013. Edinburgh SLT and MT System Description for the IWSLT 2013 Evaluation. In *Proceedings of the 10th International Workshop on Spoken Language Translation*, pages 40–48, Heidelberg, Germany, December.

Maximilian Bisani and Hermann Ney. 2004. Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 409–412, Montréal, Canada, May.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan, June.

Nadir Durrani, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013a. Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT? In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August.

Nadir Durrani, Barry Haddow, Kenneth Heafield, and Philipp Koehn. 2013b. Edinburgh's Machine Translation Systems for European Language Pairs. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August.

Nadir Durrani, Helmut Schmid, Alexander Fraser, Hassan Sajjad, and Richard Farkas. 2013c. Munich-Edinburgh-Stuttgart Submissions of OSM Systems at WMT13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria.

Nadir Durrani, Barry Haddow, Philipp Koehn, and Kenneth Heafield. 2014. Edinburgh's Phrase-based Machine Translation Systems for WMT-14. In *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, June.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *EMNLP*, pages 53–61.

M. Freitag, S. Peitz, J. Wuebker, H. Ney, N. Durrani, M. Huck, P. Koehn, T.-L. Ha, J. Niehues, M. Mediani, T. Herrmann, A. Waibel, N. Bertoldi, M. Cettolo, and M. Federico. 2013. EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project. In *International Workshop on Spoken Language Translation*, Heidelberg, Germany, December.

Markus Freitag, Matthias Huck, and Hermann Ney. 2014. Jane: Open Source Machine Translation System Combination. In *Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, April.

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, HI, USA, October.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, pages 273–280, Boston, MA, USA, May.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable Inference and Training of Context-Rich Syntactic Translation Models. In *Proc. of the 21st International Conf. on Computational Linguistics and 44th Annual Meeting of the Assoc. for Computational Linguistics*, pages 961–968, Sydney, Australia, July.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2012. Sparse Lexicalised features and Topic Adaptation for SMT. In *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 268–275.

Xiaodong He and Li Deng. 2012. Maximum Expected BLEU Training of Phrase and Lexicon Translation Models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 292–301, Jeju, Republic of Korea, July.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, UK, July.

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*, Atlanta, GA, USA, June.

Teresa Herrmann, Mohammed Mediani, Eunah Cho, Thanh-Le Ha, Jan Niehues, Isabel Slawik, Yuqi Zhang, and Alex Waibel. 2014. The Karlsruhe Institute of Technology Translation Systems for the WMT 2014. In *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, June.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A Unified Framework for Phrase-Based, Hierarchical, and Syntax-Based Statistical Machine Translation. pages 152–159, Tokyo, Japan, December.

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

Matthias Huck, Joern Wuebker, Felix Rietig, and Hermann Ney. 2013. A Phrase Orientation Model for Hierarchical Machine Translation. In *ACL 2013 Eighth Workshop on Statistical Machine Translation*, pages 452–463, Sofia, Bulgaria, August.

Matthias Huck, Hieu Hoang, and Philipp Koehn. 2014. Augmenting String-to-Tree and Tree-to-String Translation with Non-Syntactic Phrases. In *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, June.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of ACL 2003*.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *EMNLP-CoNLL*, pages 868–876, Prague, Czech Republic, June.

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

Philipp Koehn, Amittai Axelrod, Alexandra B. Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Pittsburgh, PA, USA.

P. Koehn, H. Hoang, A. Birch, 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*, pages 177–180, Prague, Czech Republic, June.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-Risk Decoding for Statistical Machine Translation. In *Proc. Human Language Technology Conf. / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL)*, pages 169–176, Boston, MA, USA, May.

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, CA, USA, December.

Arne Mauser, Saša Hasan, and Hermann 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, Singapore, August.

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 Eight International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, USA.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden, July.

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 *Proceedings of Third ACL Workshop on Statistical Machine Translation*, Columbus, USA.

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.

Franz Josef Och. 1999. An Efficient Method for Determining Bilingual Word Classes. In *EACL'99*.

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.

Stephan Peitz, Joern Wuebker, Markus Freitag, and Hermann Ney. 2014. The RWTH Aachen German-English Machine Translation System for WMT 2014. In *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, June.

Slav Petrov and Dan Klein. 2007. Improved Inference for Unlexicalized Parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April.

Slav Petrov and Dan Klein. 2008. Parsing German with Latent Variable Grammars. In *Proceedings of the Workshop on Parsing German at ACL '08*, pages 33–39, Columbus, OH, USA, June.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Assoc. for Computational Linguistics*, pages 433–440, Sydney, Australia, July.

Anna N. Rafferty and Christopher D. Manning. 2008a. Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In *Proceedings of the Workshop on Parsing German at ACL '08*, pages 40–46, Columbus, OH, USA, June.

Anna N. Rafferty and Christopher D. Manning. 2008b. 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 *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Skövde, Sweden.

Helmut Schmid and Florian Laws. 2008. Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. In *COLING 2008*, Manchester, UK.

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

Helmut Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proc. of the Int. Conf. on Computational Linguistics (COLING)*, Geneva, Switzerland, August.

Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2013*, pages 601–609, Hissar, Bulgaria.

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, CO, USA, September.

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, Michigan, USA.

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.

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

Philip Williams and Philipp Koehn. 2012. GHKM Rule Extraction and Scope-3 Parsing in Moses. In *Proceedings of the Seventh Workshop on Statistical Machine Translation (WMT)*, pages 388–394, Montréal, Canada, June.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, Eva Hasler, and Philipp Koehn. 2014. Edinburgh's Syntax-Based Systems at WMT 2014. In *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, June.

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 *COLING '12: The 24th Int. Conf. on Computational Linguistics*, pages 483–491, Mumbai, India, December.

Joern Wuebker, Stephan Peitz, Felix Rietig, and Hermann Ney. 2013. Improving Statistical Machine Translation with Word Class Models. In *Conference on Empirical Methods in Natural Language Processing*, pages 1377–1381, Seattle, WA, USA, October.