Combined Spoken Language Translation

[*]Markus Freitag, [*]Joern Wuebker, [*]Stephan Peitz, [*]Hermann Ney,
[‡]Matthias Huck, [‡]Alexandra Birch, [‡]Nadir Durrani, [‡]Philipp Koehn,
[†]Mohammed Mediani, [†]Isabel Slawik, [†]Jan Niehues, [†]Eunah Cho, [†]Alex Waibel,
[§]Nicola Bertoldi, [§]Mauro Cettolo, [§]Marcello Federico
[*]RWTH Aachen University, Aachen, Germany
[‡]University of Edinburgh, Edinburgh, Scotland
[†]Karlsruhe Institute of Technology, Karlsruhe, Germany
[§]Fondazione Bruno Kessler, Trento, Italy
[*]{freitag,wuebker,peitz,ney}@cs.rwth-aachen.de
[‡]a.birch@ed.ac.uk {mhuck,dnadir,pkoehn}@inf.ed.ac.uk
[†]{firstname.lastname}@kit.edu
[§]{bertoldi,cettolo,federico}@fbk.eu

Abstract

EU-BRIDGE[1] is a European research project which is aimed at developing innovative speech translation technology. One of the collaborative efforts within EU-BRIDGE is to produce joint submissions of up to four different partners to the evaluation campaign at the 2014 International Workshop on Spoken Language Translation (IWSLT). We submitted combined translations to the German→English spoken language translation (SLT) track as well as to the German→English, English→German and English→French machine translation (MT) tracks. In this paper, we present the techniques which were applied by the different individual translation systems of RWTH Aachen University, the University of Edinburgh, Karlsruhe Institute of Technology, and Fondazione Bruno Kessler. We then show the combination approach developed at RWTH Aachen University which combined the individual systems. The consensus translations yield empirical gains of up to 2.3 points in BLEU and 1.2 points in TER compared to the best individual system.

1. Introduction

The EU-BRIDGE project is funded by the European Union under the Seventh Framework Programme (FP7) and brings together several project partners who have each previously been very successful in contributing to advancements in automatic speech recognition and statistical machine translation. A number of languages and language pairs (both well-covered and under-resourced ones) are tackled with automatic speech recognition (ASR) and MT technology with different use cases in mind. Four of the EU-BRIDGE project partners are particularly experienced in machine translation for European language pairs: RWTH Aachen University (RWTH), the University of Edinburgh (UEDIN), Karlsruhe Institute of Technology (KIT), and Fondazione Bruno Kessler (FBK) have all regularly participated in large-scale evaluation campaigns like IWSLT and WMT in recent years, thereby demonstrating their ability to continuously enhance their systems and promoting progress in machine translation. Machine translation research within EU-BRIDGE has a strong focus on translation of spoken language. The IWSLT TED talks task constitutes an interesting framework for empirical testing of some of the systems for spoken language translation which are developed as part of the project.

In this work, we describe the EU-BRIDGE submissions to the 2014 IWSLT translation task. This year, we combined several single systems of RWTH, UEDIN, KIT, and FBK for the German→English SLT, German→English MT, English→German MT, and English→French MT tasks. Additionally to the standard system combination pipeline presented in [1, 2], we applied a recurrent neural network rescoring step [3] for the English→French MT task. Similar cooperative approaches based on system combination have proven to be valuable for machine translation in previous joint submissions, e.g. [4, 5].

2. RWTH Aachen University

RWTH applied the identical training pipeline and models on both language pairs: The state-of-the-art phrase-based baseline systems were augmented with a hierarchical reordering model, several additional language models (LMs) and maximum expected BLEU training for phrasal, lexical and reordering models. Further, RWTH employed rescoring with novel recurrent neural language and translation models. The same systems were used for the SLT track, where RWTH ad-

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

ditionally performed punctuation prediction on the automatic transcriptions employing hierarchical phrase-based translation. Both the phrase-based and the hierarchical decoder are implemented in RWTH's publicly available translation toolkit Jane [6, 7]. The model weights of all systems were tuned with standard Minimum Error Rate Training [8] on the provided dev2012 set. RWTH used BLEU as optimization objective. For the German→English translation direction, in a preprocessing step the German source was decompounded [9] and part-of-speech-based long-range verb reordering rules [10] were applied. RWTH's translation systems are described in more detail in [11].

Backoff Language Models
Each translation system used three backoff LMs that were estimated with the KenLM toolkit [12]: A large general domain 5-gram LM, an in-domain 5-gram LM and a 7-gram word class language model (wcLM). All of them used interpolated Kneser-Ney smoothing. For the general domain LM, RWTH first selected $\frac{1}{2}$ of the English Shuffled News, and $\frac{1}{4}$ of the French Shuffled News as well as both the English and French Gigaword corpora by the cross-entropy difference criterion described in [13]. The selection was then concatenated with all available remaining monolingual data and used to build and unpruned LM. The in-domain language models were estimated on the TED data only. For the word class LM, RWTH trained 200 classes on the target side of the bilingual training data using an in-house tool similar to `mkcls` [14]. With these class definitions, RWTH applied the technique shown in [15] to compute the wcLM on the same data as the general-domain LM.

Maximum Expected BLEU Training
RWTH applied discriminative training, learning three types of features under a maximum expected BLEU objective [16]. It was performed on the TED portion of the data, which is high quality in-domain data of reasonable size. This makes training feasible while at the same time providing an implicit domain adaptation effect. Similar to [16], RWTH generated 100-best lists on the training data which were used as training samples for a gradient based update method. Leave-one-out [17] was applied to circumvent over-fitting. Here, RWTH followed an approach similar to [18], where each feature type was condensed into a single feature for the log-linear model combination. In the first pass, RWTH trained phrase pair and phrase-internal word pair features, and in the second pass a hierarchical reordering model, resulting altogether in an additional eight models for log-linear combination.

Recurrent Neural Network Models
All systems applied rescoring on 1000-best lists using recurrent language and translation models. The recurrency was handled with the long short-term memory (LSTM) architecture [19] and RWTH used a class-factored output layer for increased efficiency as described in [20]. All neural networks were trained on the TED portion of the data with 2000 word classes. In addition to the recurrent language model (RNN-LM), RWTH applied the deep bidirectional word-based translation model (RNN-BTM) described in [3], which is capable of taking the *full source context* into account for each translation decision.

Spoken Language Translation
For the SLT task, RWTH reintroduced punctuation and case information before the actual translation similar to [21]. However, RWTH employed a hierarchical phrase-based system with a maximum of one nonterminal symbol per rule in place of a phrase-based system. A punctuation prediction system based on hierarchical translation is able to capture long-range dependencies between words and punctuation marks and is more robust for unseen word sequences. The model weights are tuned with standard MERT on 100-best lists. As optimization criterion RWTH used F_2-Score rather than BLEU or WER. More details can be found in [22].

Since punctuation predicting and recasing were applied before the actual translation, the final translation systems from the MT track could be kept completely unchanged.

3. University of Edinburgh

The UEDIN translation engines [23] are based on the open source Moses toolkit [24]. UEDIN set up phrase-based systems for all SLT and MT tasks covered in this paper, and additionally a string-to-tree syntax-based system [25] for the English→German MT task. The systems were trained using monolingual and parallel data from WIT^3, Europarl, MultiUN, the English and French Gigaword corpora as provided by the Linguistic Data Consortium, the German Political Speeches Corpus, and the Common Crawl, 10^9, and News Commentary corpora from the WMT shared task training data. Word alignments for the MT track systems were created by aligning the data in both directions with MGIZA++ [26] and symmetrizing the two trained alignments. Word alignments for the SLT track system were created using fast_align [27]. The SRILM toolkit [28] was employed to train 5-gram LMs with modified Kneser-Ney smoothing [29]. UEDIN trained individual LMs on each corpus and then interpolated them using weights tuned to minimize perplexity on a development set.

Common features included in the UEDIN phrase-based systems are the language model, phrase translation scores in both directions smoothed with Good-Turing discounting, lexical translation scores in both directions, word and phrase penalties, six simple count-based binary features, distance-based distortion costs, a hierarchical lexicalized reordering model [30], sparse lexical and domain indicator features [31] and operation sequence models over different word representations [32]. Model weights were optimized with batch MIRA [33] to maximize BLEU [34].

Spoken Language Translation
One of the main challenges of spoken language translation is to overcome the mismatch in the style of data that the

speech recognition systems output, and the written text that is used to train the translation model. ASR system output lacks punctuation and capitalization, which is the main stylistic differences. Previous research [35, 21, 36] suggests that it is preferrable to punctuate the text before translation, which is what UEDIN did by training a translation system on the German side of the parallel data. The "source language" of the system had punctuation and capitalization stripped, and the "target language" was the standard German parallel text. The handling of punctuation is similar to the other groups in this paper, however UEDIN used a phrase-based model with no distortion or reordering, and tuned the model to the ASR input text using batch MIRA and the BLEU score.

German→English MT
For the UEDIN German→English MT task system, pre-reordering [37] and compound splitting [38] were applied to the German source language side in a preprocessing step. A factored translation model [39] was employed. Source side factors are word, lemma, part-of-speech (POS) tag, and morphological tag. Target side factors are word, lemma, and POS tag. UEDIN incorporated two additional LMs into the German→English MT system: a 7-gram LM over POS tags (trained on WIT^3 only) and a 7-gram LM over lemmas (trained on WIT^3 only). Model weights were optimized on a concatenation of dev2010 and dev2012.

English→French MT
UEDIN contributed two phrase-based systems for the English→French EU-BRIDGE system combination. Both comprise Brown clusters with 200 classes as additional factors on source and target side. The system denoted as UEDIN-A was trained without the MultiUN and 10^9 corpora, the system denoted as UEDIN-B was trained with all corpora. An additional feature incorporated into the systems is an LM over Brown clusters (UEDIN-A: 7-gram, UEDIN-B: 5-gram). Model weights were optimized on dev2010.

English→German MT
UEDIN contributed two phrase-based systems (UEDIN-A and UEDIN-B) and a syntax-based system (UEDIN-C) for English→German MT.

Phrase-based systems. UEDIN-A and UEDIN-B employ factored models. Source side factors are word, POS tag, and Brown cluster (2000 classes). Target side factors are word, POS tag, Brown cluster (2000 classes), and morphological tag. UEDIN-A was trained with all corpora, whereas for UEDIN-B the parallel training data was restricted to the in-domain WIT^3 corpus. Additional features of the systems are: a 5-gram LM over Brown clusters, a 7-gram LM over morphological tags (UEDIN-A: trained on all data, UEDIN-B: trained on WIT^3 only), and a 7-gram LM over POS tags (UEDIN-A, not UEDIN-B). Model weights of UEDIN-B were optimized on dev2010, model weights of UEDIN-A on a concatenation of dev2010 and dev2012.

Syntax-based system. UEDIN-C is a string-to-tree translation system with similar features as the ones described in [40]. The target-side data was parsed with BitPar [41], and right binarization was applied to the parse trees. The system was adapted to the TED domain by extracting separate rule tables (from the WIT^3 corpus and from the rest of the parallel data) and merging them with a fill-up technique [42]. Augmenting the system with non-syntactic phrases [43] and adding soft source syntactic constraints [44] yielded further improvements. Model weights of UEDIN-C were optimized on a concatenation of dev2010 and dev2012.

4. Karlsruhe Institute of Technology

The KIT translations were generated by an in-house phrase-based translations system [45]. The models were trained on the Europarl, News Commentary, WIT^3, Common Crawl corpora for all directions, as well as on the additional monolingual training data. The noisy Crawl corpora were filtered using an SVM classifier [46]. In addition to the standard preprocessing, KIT used compound splitting [38] for the German text when translating from German. In the SLT task, KIT first recased the input and added punctuation marks to the ASR hypotheses. This was done with a monolingual translation system as shown in [36].

In all translation directions, KIT used a pre-reordering approach. Different reorderings of the source sentences were encoded in a word lattice. For the English→French system, only short-range rules were used to generate these lattices [47]. Long-range rules [48] and tree-based reordering rules [49] were used for German→English. The POS tags needed for these rules were generated by the TreeTagger [50] and the parse trees by the Stanford Parser [51]. In addition, for the language pairs involving German KIT applied the different reorderings of both language pairs using a lexicalized reordering model. The phrase tables of the systems were trained using GIZA++ alignment [52]. KIT adapted the phrase table to the TED domain using the backoff approach and by means of candidate selection [53]. In addition to the phrase table probabilities, KIT modeled the translation process by a bilingual language model [54] and a discriminative word lexicon using source context features [55].

During decoding, KIT used several LMs to adapt the system to the task and to better model the sentence structure using a class-based LM. For the German→English task, KIT used one LM trained on all data, an in-domain LM trained only on the WIT^3 corpus, and one LM trained on 5M sentences selected using cross-entropy difference [13]. As classes KIT used the clusters obtained using the `mkcls` algorithm on the WIT^3 corpus. For German↔ English, KIT used a 9-gram LM with 100 or 1000 clusters and for the English→French MT task, a cluster-based 4-gram LM was trained on 500 clusters. For English→German, KIT also used a 9-gram POS-based LM. The log-linear combination of all these models was optimized on the provided development data using MERT.

5. Fondazione Bruno Kessler

The FBK system was built upon a standard phrase-based system using the Moses toolkit [24], and exploited the huge amount of parallel English-French and monolingual French training data provided by the organizers. It featured a statistical log-linear model including a phrase-based translation model (TM) and lexicalized phrase-based reordering models (RM), two French language models (LMs), as well as distortion, word and phrase penalties. Tuning of the system was performed on dev2010 by optimizing BLEU using Minimum Error Rate Training [8]. It is worth noticing that all available development data sets, namely dev2010 and test2010-2012, were added to the in-domain training data to build the system actually employed for the 2014 evaluation campaign.

In order to adapt the system on TED specific domain and genre and to reduce the size of the system, data selection was carried out on all parallel English-French corpora, using the whole WIT^3 [56] training corpus as in-domain data. Data selection was performed by means of XenC toolkit [57] exploiting bilingual cross-entropy difference [58] separately for each available training corpus except the in-domain WIT^3 data. Different amount of texts were selected from each corpora ranging from 2% to 30%, and then concatenating for building one parallel corpus containing 2.6M sentences for a total of 57M English and 63M French running words.

Two TMs and two RMs were trained independently on the parallel in-domain and selected data, using the standard Moses procedure and MGIZA++ toolkit [26] for word-alignment; TMs and RMs were combined using the back-off technique (for both TM and RM), taking WIT^3 as primary component, for a total of 168M phrase pairs. The back-off table combination is similar to the fill-up technique [42], but does not add any provenance binary features.

The French side of in-domain and selected data were also employed to estimate a 2-component mixture language model [59]. Moreover, a second huge French LM was estimated on all permitted monolingual French data consisting of $\sim$1.4G running words, as a mixture of 8 components. Both LMs have order 5 and were smoothed by means of the interpolated Improved Kneser-Ney method [29]; they include 57M and 661M 5-grams, respectively. A full description of the system can be found in the FBK system paper.

6. System Combination

In this section, we give a brief re-introduction of confusion network 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 applied a system combination implementation that has been developed at RWTH Aachen University [1].

In Fig. 1 an overview is illustrated. We first address the generation of a confusion network (CN) from I input translations. For that we need a pairwise alignment between all input hypotheses. This alignment is calculated via METEOR [60]. The hypotheses are then reordered to match the word order of a selected skeleton hypothesis. Instead of using only one of the input hypothesis as skeleton, we generate I different CNs, each having one of the input systems as skeleton. The final lattice is the union of all I previous generated CNs. In Fig. 2 an example confusion network of $I = 4$ input translations with one skeleton translation is illustrated. Between two adjacent nodes, we always have a choice between the I different system output words. The confusion network decoding step involves determining the shortest path through the network. Each arc is assigned one score which is a linear model combination (Eq. 1) of M different models.

$$\sum_{m=1}^{M} \lambda_m h_m \tag{1}$$

The standard set of models is a word penalty, a 3-gram language model trained on the input hypotheses, and for each system one binary voting feature. During decoding the binary voting feature for system i ($1 \leq i \leq I$) is 1 iff the word is from system i, otherwise 0. The M different model weights λ_m are trained with MERT [8].

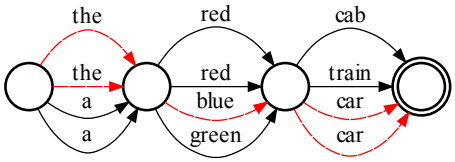

Figure 2: System A: *the red cab* ; System B: *the red train* ; System C: *a blue car* ; System D: *a green car* ; Reference: *the blue car* .

7. Results

In this section, we present our experimental results. All reported BLEU [34] and TER [61] scores are case-sensitive with one reference. All system combination results have been generated with RWTH's open source system combination implementation Jane [1].

German→English SLT

For the German→English SLT task, we combined three different individual systems generated by UEDIN, KIT, and RWTH. Experimental results are given in Table 1. The final system combination yields improvements of 1.5 points in BLEU and 1.2 points in TER compared to the best single system (KIT). All single systems as well as the system combination parameters were tuned on dev2012. For this year's IWSLT SLT track, dev2012 was the only given test set containing automatic speech recognition output.

German→English MT

Similar to the SLT track, the German→English MT system combination submission is a combined translation of three different individual systems by UEDIN, KIT, and RWTH.

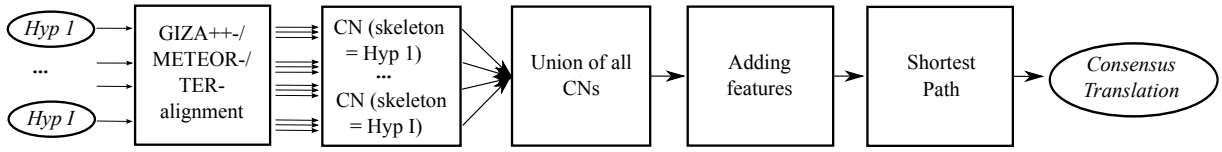

Figure 1: Confusion network decoding structure.

Table 1: Results for the German→English SLT task.

system	dev2012	
	BLEU	TER
KIT	20.7	60.5
RWTH	20.8	61.4
UEDIN	20.3	63.0
syscom	22.2	59.3

Table 2: Results for the German→English MT task.

system	tst2010		tst2011		tst2012	
	BLEU	TER	BLEU	TER	BLEU	TER
KIT	31.5	47.6	37.1	42.5	32.0	47.6
RWTH	31.8	47.2	38.3	41.3	32.0	47.0
UEDIN	31.6	47.6	37.3	42.5	31.7	47.9
syscom	33.3	46.1	39.4	40.6	33.5	46.2

Experimental results are given in Table 2. The system combination parameters have been optimized on test2012. Compared to the best individual system (RWTH), the system combination improved translation scores by up to 1.5 points in BLEU and 1.1 points in TER.

English→French MT

For the English→French MT task, we combined five different individual systems. FBK, KIT, and RWTH provided one individual system output for the system combination. UEDIN added one advanced contrastive system in addition to their primary system. Experimental results are given in Table 3. The system combination of all five individual systems yields an improvement of up to 0.6 points in BLEU compared to the best RWTH individual system output. Using a recurrent neural network (RNN) LM to rescore a 1000-best list of the system combination output, leads to a small translation improvement of +0.1 in BLEU. The same RNN LM was applied in the best individual system of RWTH Aachen. The improvements are only small, as the model is already contained the best individual system.

English→German MT

For the English→German setup, we combined three different individual system setups of UEDIN with the primary submission of KIT. Experimental results are given in Table 4. All system combination parameters are tuned on tst2012. The EU-BRIDGE submission enhanced the translation quality by up to 1.4 points in BLEU and 1.2 points in TER compared to the best individual system.

Table 3: Results for the English→French MT task.

system	tst2010		tst2011		tst2012	
	BLEU	TER	BLEU	TER	BLEU	TER
FBK	32.8	50.4	39.2	42.6	40.0	41.4
KIT	33.1	48.4	37.3	42.5	39.1	40.2
RWTH	34.5	47.6	41.1	40.1	42.0	38.6
UEDIN-A	33.6	48.5	40.2	40.6	41.0	39.6
UEDIN-B	33.2	49.1	39.1	42.0	40.7	39.8
syscom	35.1	48.5	41.7	41.4	44.0	38.7
+RNN	35.2	48.5	41.7	41.3	44.3	38.5

Table 4: Results for the English→German MT task.

system	tst2010		tst2011		tst2012	
	BLEU	TER	BLEU	TER	BLEU	TER
KIT	24.5	55.2	27.1	50.5	23.5	56.0
UEDIN-A	24.9	55.5	27.8	50.1	23.4	56.9
UEDIN-B	24.1	55.7	26.7	50.8	22.2	57.3
UEDIN-C	24.8	55.3	26.5	50.5	23.1	56.6
syscom	25.9	54.0	28.1	49.1	24.9	55.0

8. Conclusion

We achieved better translation performance with gains of up to +2.3 points in BLEU and -1.2 points in TER by combining the different system hypotheses of up to four partners of the EU-BRIDGE project. The four research institutes (RWTH Aachen University, University of Edinburgh, Karlsruhe Institute of Technology, Fondazione Bruno Kessler) are maintaining different machine translation engines based on different approaches. System combination combined all the different advancements of all engines together into our final submissions. For English→French we applied a recurrent neural network language model in an additional rescoring step which only gives small improvement of +0.1 points in BLEU.

9. 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.

10. References

[1] M. Freitag, M. Huck, and H. Ney, "Jane: Open Source Machine Translation System Combination," in *Proc. of the Conf. of the European Chapter of the Assoc. for Computational Linguistics (EACL)*, Gothenburg, Sweden, Apr. 2014, pp. 29–32.

[2] E. Matusov, N. Ueffing, and H. Ney, "Computing Con-

sensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment," in *Proc. of the Conf. of the European Chapter of the Assoc. for Computational Linguistics (EACL)*, Trento, Italy, Apr. 2006, pp. 33–40.

[3] M. Sundermeyer, T. Alkhouli, J. Wuebker, and H. Ney, "Translation modeling with bidirectional recurrent neural networks," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014.

[4] M. Freitag, S. Peitz, J. Wuebker, H. Ney, M. Huck, R. Sennrich, N. Durrani, M. Nadejde, P. Williams, P. Koehn, T. Herrmann, E. Cho, and A. Waibel, "EU-BRIDGE MT: Combined Machine Translation," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Baltimore, MD, USA, June 2014, pp. 105–113.

[5] 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, "EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, Dec. 2013, pp. 128–135.

[6] D. Vilar, D. Stein, M. Huck, and H. Ney, "Jane: Open source hierarchical translation, extended with reordering and lexicon models," in *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, Uppsala, Sweden, July 2010, pp. 262–270.

[7] J. Wuebker, M. Huck, S. Peitz, M. Nuhn, M. Freitag, J.-T. Peter, S. Mansour, and H. Ney, "Jane 2: Open Source Phrase-based and Hierarchical Statistical Machine Translation," in *COLING '12: The 24th Int. Conf. on Computational Linguistics*, Mumbai, India, Dec. 2012, pp. 483–491.

[8] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Sapporo, Japan, July 2003, pp. 160–167.

[9] P. Koehn and K. Knight, "Empirical Methods for Compound Splitting," in *Proceedings of European Chapter of the ACL (EACL 2009)*, 2003, pp. 187–194.

[10] M. Popović and H. Ney, "POS-based Word Reorderings for Statistical Machine Translation," in *International Conference on Language Resources and Evaluation*, 2006, pp. 1278–1283.

[11] J. Wuebker, S. Peitz, A. Guta, and H. Ney, "The RWTH Aachen Machine Translation Systems for IWSLT 2014," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, South Lake Tahoe, CA, USA, Dec. 2014.

[12] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified Kneser-Ney language model estimation," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Sofia, Bulgaria, August 2013, pp. 690–696.

[13] R. Moore and W. Lewis, "Intelligent selection of language model training data," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Uppsala, Sweden, July 2010, pp. 220–224.

[14] F. J. Och, "An Efficient Method for Determining Bilingual Word Classes." in *Proc. of the Conf. of the European Chapter of the Assoc. for Computational Linguistics (EACL)*, 1999.

[15] J. Wuebker, S. Peitz, F. Rietig, and H. Ney, "Improving Statistical Machine Translation with Word Class Models," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Seattle, WA, USA, Oct. 2013, pp. 1377–1381.

[16] X. He and L. Deng, "Maximum Expected BLEU Training of Phrase and Lexicon Translation Models," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Jeju, Republic of Korea, July 2012, pp. 292–301.

[17] J. Wuebker, A. Mauser, and H. Ney, "Training Phrase Translation Models with Leaving-One-Out," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Uppsala, Sweden, July 2010, pp. 475–484.

[18] M. Auli, M. Galley, and J. Gao, "Large Scale Expected BLEU Training of Phrase-based Reordering Models," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[20] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Interspeech*, Portland, OR, USA, Sept. 2012.

[21] S. Peitz, M. Freitag, A. Mauser, and H. Ney, "Modeling Punctuation Prediction as Machine Translation," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, USA, Dec. 2011.

[22] S. Peitz, M. Freitag, and H. Ney, "Better Punctuation Prediction with Hierarchical Phrase-Based Translation," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, South Lake Tahoe, CA, USA, Dec. 2014.

[23] A. Birch, M. Huck, N. Durrani, N. Bogoychev, and P. Koehn, "Edinburgh SLT and MT System Description for the IWSLT 2014 Evaluation," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, South Lake Tahoe, CA, USA, Dec. 2014.

[24] 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, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Prague, Czech Republic, June 2007, pp. 177–180.

[25] M. Galley, M. Hopkins, K. Knight, and D. Marcu, "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)*, Boston, MA, USA, May 2004, pp. 273–280.

[26] Q. Gao and S. Vogel, "Parallel Implementations of Word Alignment Tool," in *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, ser. SETQA-NLP '08, Columbus, OH, USA, June 2008, pp. 49–57.

[27] C. Dyer, V. Chahuneau, and N. A. Smith, "A Simple, Fast, and Effective Reparameterization of IBM Model 2," in *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, Atlanta, GA, USA, June 2013, pp. 644–648.

[28] A. Stolcke, "SRILM – an Extensible Language Modeling Toolkit," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, Denver, CO, USA, Sept. 2002, pp. 901–904.

[29] S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," Computer Science Group, Harvard University, Cambridge, MA, USA, Tech. Rep. TR-10-98, Aug. 1998.

[30] M. Galley and C. D. Manning, "A Simple and Effective Hierarchical Phrase Reordering Model," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Honolulu, HI, USA, Oct. 2008, pp. 847–855.

[31] E. Hasler, B. Haddow, and P. Koehn, "Sparse Lexicalised Features and Topic Adaptation for SMT," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, Dec. 2012, pp. 268–275.

[32] N. Durrani, P. Koehn, H. Schmid, and A. Fraser, "Investigating the Usefulness of Generalized Word Representations in SMT," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland, Aug. 2014, pp. 421–432.

[33] C. Cherry and G. Foster, "Batch Tuning Strategies for Statistical Machine Translation," in *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, Montréal, Canada, June 2012, pp. 427–436.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Philadelphia, PA, USA, July 2002, pp. 311–318.

[35] E. Matusov, A. Mauser, and H. Ney, "Automatic sentence segmentation and punctuation prediction for spoken language translation," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, November 2006.

[36] E. Cho, J. Niehues, and A. Waibel, "Segmentation and Punctuation Prediction in Speech Language Translation using a Monolingual Translation System," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[37] M. Collins, P. Koehn, and I. Kucerova, "Clause Restructuring for Statistical Machine Translation," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Ann Arbor, MI, USA, June 2005, pp. 531–540.

[38] P. Koehn and K. Knight, "Empirical Methods for Compound Splitting," in *Proc. of the Conf. of the European Chapter of the Assoc. for Computational Linguistics (EACL)*, Budapest, Hungary, Apr. 2003, pp. 187–194.

[39] P. Koehn and B. Haddow, "Interpolated Backoff for Factored Translation Models," in *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, San Diego, CA, USA, Oct./Nov. 2012.

[40] P. Williams, R. Sennrich, M. Nadejde, M. Huck, E. Hasler, and P. Koehn, "Edinburgh's Syntax-Based Systems at WMT 2014," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Baltimore, MD, USA, June 2014, pp. 207–214.

[41] H. Schmid, "Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors," in *Proc. of the Int. Conf. on Computational Linguistics (COLING)*, Geneva, Switzerland, Aug. 2004.

[42] A. Bisazza, N. Ruiz, and M. Federico, "Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, USA, Dec. 2011, pp. 136–143.

[43] M. Huck, H. Hoang, and P. Koehn, "Augmenting String-to-Tree and Tree-to-String Translation with Non-Syntactic Phrases," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Baltimore, MD, USA, June 2014, pp. 486–498.

[44] Huck, Matthias and Hoang, Hieu and Koehn, Philipp, "Preference Grammars and Soft Syntactic Constraints for GHKM Syntax-based Statistical Machine Translation," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar, Oct. 2014, pp. 148–156.

[45] S. Vogel, "SMT Decoder Dissected: Word Reordering." in *International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China, 2003.

[46] M. Mediani, E. Cho, J. Niehues, T. Herrmann, and A. Waibel, "The KIT English-French Translation systems for IWSLT 2011," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, USA, 2011.

[47] K. Rottmann and S. Vogel, "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, 2007.

[48] J. Niehues, T. Herrmann, M. Kolss, and A. Waibel, "The Universität Karlsruhe Translation System for the EACL-WMT 2009," in *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT 2009)*, Athens, Greece, 2009.

[49] T. Herrmann, J. Niehues, and A. Waibel, "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, GA, USA, June 2013.

[50] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," in *Int. Conf. on New Methods in Language Processing*, Manchester, UK, 1994.

[51] A. N. Rafferty and C. D. Manning, "Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines," in *Proc. of the Workshop on Parsing German*, Columbus, OH, USA, 2008.

[52] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, Mar. 2003.

[53] J. Niehues and A. Waibel, "Detailed Analysis of Different Strategies for Phrase Table Adaptation in SMT," in *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, 2012.

[54] J. Niehues, T. Herrmann, S. Vogel, and A. Waibel, "Wider Context by Using Bilingual Language Models in Machine Translation," in *Sixth Workshop on Statistical Machine Translation (WMT 2011)*, Edinburgh, UK, 2011.

[55] J. Niehues and A. Waibel, "An MT Error-Driven Discriminative Word Lexicon using Sentence Structure Features," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, 2013, pp. 512–520.

[56] M. Cettolo, C. Girardi, and M. Federico, "WIT^3: Web Inventory of Transcribed and Translated Talks," in *Proc. of the Annual Conf. of the European Assoc. for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[57] A. Rousseau, "Xenc: An open-source tool for data selection in natural language processing," *The Prague Bulletin of Mathematical Linguistics*, no. 100, pp. 73–82, 2013.

[58] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, 2011, pp. 355–362.

[59] M. Federico and R. De Mori, "Language modelling," in *Spoken Dialogues with Computers*, R. D. Mori, Ed. London, UK: Academy Press, 1998, ch. 7, pp. 199–230.

[60] S. Banerjee and A. Lavie, "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*, Ann Arbor, MI, USA, June 2005, pp. 65–72.

[61] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, Cambridge, MA, USA, Aug. 2006, pp. 223–231.