EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project

[*]Markus Freitag, [*]Stephan Peitz, [*]Joern Wuebker, [*]Hermann Ney,
[‡]Nadir Durrani, [‡]Matthias Huck, [‡]Philipp Koehn,
[†]Thanh-Le Ha, [†]Jan Niehues, [†]Mohammed Mediani, [†]Teresa Herrmann, [†]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,peitz,wuebker,ney}@cs.rwth-aachen.de
[‡]{ndurrani,mhuck,pkoehn}@inf.ed.ac.uk
[†]{thanh-le.ha,jan.niehues,teresa.herrmann,alex.waibel}@kit.edu
[†]mmediani@ira.uka.de
[§]{bertoldi,cettolo,federico}@fbk.eu

Abstract

EU-BRIDGE[1] is a European research project which is aimed at developing innovative speech translation technology. 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, English→French and German→English. Four research institutions involved in the EU-BRIDGE project combined their individual machine translation systems and participated with a joint setup in the machine translation track of the evaluation campaign at the 2013 International Workshop on Spoken Language Translation (IWSLT).

We present the methods and techniques to achieve high translation quality for text translation of talks which are applied at RWTH Aachen University, the University of Edinburgh, Karlsruhe Institute of Technology, and Fondazione Bruno Kessler. We then show how we have been able to considerably boost translation performance (as measured in terms of the metrics BLEU and TER) by means of system combination. The joint setups yield empirical gains of up to 1.4 points in BLEU and 2.8 points in TER on the IWSLT test sets compared to the best single systems.

1. Introduction

The International Workshop on Spoken Language Translation [1] hosts a yearly open evaluation campaign on the translation of TED talks [2]. The TED talks task is challenging from the perspective of automatic speech recognition (ASR) and machine translation (MT) as it involves spontaneous speech and heterogeneous topics and styles. The task is open domain, with a wide range of heavily dissimilar subjects and jargons across talks. IWSLT subdivides the task and separately evaluates *automatic transcription of talks from audio to text*, *speech translation of talks from audio*, and *text translation of talks* as three different tracks [3, 4]. The training data is constrained to the corpora specified by the organizers. The supplied list of corpora comprises a large amount of publicly available monolingual and parallel training data, though, including WIT[3] [5], Europarl [6], MultiUN [7], the English and French Gigaword corpora as provided by the Linguistic Data Consortium [8], and the News Crawl, 10^9 and News Commentary corpora from the WMT shared task training data [9]. For the two "official" language pairs [1] for translation at IWSLT 2013, English→French and German→English, these resources allow for building of systems with state-of-the-art performance by participants.

The EU-BRIDGE project is funded by the European Union under the Seventh Framework Programme (FP7) [10] 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 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 trans-

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

lation 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.

The work described here is an attempt to attain translation quality beyond strong single system performance via system combination [11]. Similar cooperative approaches based on system combination have proven to be valuable for machine translation in other projects, e.g. in the Quaero programme [12, 13]. Within EU-BRIDGE, we built combined system setups for text translation of talks from English to French as well as from German to English. We found that the combined translation engines of RWTH, UEDIN, KIT, and FBK systems are very effective. In the rest of the paper we will give some insight into the technology behind the combined engines which have been used to produce the joint EU-BRIDGE submission to the IWSLT 2013 MT track.

The remainder of the paper is structured as follows: We first describe the individual English→French and German→English systems by RWTH Aachen University (Section 2), the University of Edinburgh (Section 3), Karlsruhe Institute of Technology (Section 4), and Fondazione Bruno Kessler (Section 5), respectively. We then present the techniques for machine translation system combination which have been employed to obtain consensus translations from the outputs of the individual systems of the project partners (Section 6). Experimental results in BLEU [14] and TER [15] are given in Section 7. A brief error analysis on selected examples from the test data has been conducted which we discuss in Section 8. We finally conclude the paper with Section 9.

2. RWTH Aachen University

RWTH applied both the phrase-based (*RWTH scss*) and the hierarchical (*RWTH hiero*) decoder implemented in RWTH's publicly available translation toolkit Jane [16, 17, 18, 19]. The model weights of all systems were tuned with standard Minimum Error Rate Training [20] on the provided dev2010 set. RWTH used BLEU as optimization objective. Language models were created with the SRILM toolkit [21]. All RWTH systems include the standard set of models provided by Jane.

For English→French, the final setups for *RWTH scss* and *RWTH hiero* differ in the amount of training data and in the choice of models.

For the English→French hierarchical setup the bilingual data was limited to the in-domain WIT^3 data, News Commentary, Europarl, and Common Crawl corpora. The word alignment was created with *fast_align* [22]. A language model was trained on the target side of all available bilingual data plus $\frac{1}{2}$ of the Shuffled News corpus and $\frac{1}{4}$ of the French Gigaword Second Edition corpus. The monolingual data selection for using only parts of the corpora is based on cross-entropy difference as described in [23]. The hierarchical system was extended with a second translation model. The additional translation model was trained on the WIT^3 portion of the training data only.

For the English→French phrase-based setup, RWTH utilized all available parallel data and trained a word alignment with GIZA++ [24]. The same language model as in the hierarchical setup was used. RWTH applied the following supplementary features for the phrase-based system: a lexicalized reordering model [25], a discriminative word lexicon [26], a 7-gram word class language model [27], a continuous space language model [28], and a second translation model from the WIT^3 portion of the training data only.

For German→English, RWTH decompounded the German source in a preprocessing step [29] and applied part-of-speech-based long-range verb reordering rules [30]. Both systems *RWTH scss* and *RWTH hiero* rest upon all available bilingual data and word alignment obtained with GIZA++. A language model was trained on the target side of all available bilingual data plus $\frac{1}{2}$ of the Shuffled News corpus and $\frac{1}{4}$ of the English Gigaword v3 corpus, resulting in a total of 1.7 billion running words.

In both German→English systems, RWTH applied a more sophisticated discriminative phrase training method. Similar to [31], a gradient-based method is used to optimize a maximum expected BLEU objective, for which we define BLEU on the sentence level with smoothed 3-gram and 4-gram precisions. RWTH performed discriminative training on the WIT^3 portion of the training data.

The German→English phrase-based system was furthermore improved by a lexicalized reordering model and 7-gram word class language model. RWTH finally applied domain adaptation by adding a second translation model to the decoder which was trained on the WIT^3 portion of the data only. This second translation model was likewise improved with discriminative phrase training.

3. University of Edinburgh

UEDIN's systems were trained using the Moses system [32], replicating the settings described in [33] developed for the 2013 Workshop on Statistical Machine Translation. The characteristics of the system 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 [34] used at runtime, a lexically-driven 5-gram operation sequence model [35] with four additional supportive features (two gap-based penalties, one distance-based feature and one deletion penalty), msd-bidirectional-fe lexicalized reordering, sparse lexical and domain features [36], a distortion limit of 6, 100-best translation options, minimum Bayes risk decoding [37], cube pruning [38] with a stack size of 1000 during tuning and 5000 during testing and the no-reordering-over-punctuation heuristic. UEDIN used the compact phrase table representation by [39]. For English→German, UEDIN used a sequence model over morphological tags.

The UEDIN systems were tuned on the dev2010 set made available for the IWSLT 2013 workshop. Tuning was performed using the *k*-best batch MIRA algorithm [40] with a maximum number of iterations of 25. BLEU was used as the metric to evaluate results.

While UEDIN's main submission also includes sequence models and operation sequence models over Brown word clusters, these setups were not finished in time for the contribution to the EU-BRIDGE system combination.

4. Karlsruhe Institute of Technology

The KIT translations have been generated by an in-house phrase-based translations system [41]. The models were trained on the Europarl, News Commentary, WIT^3, Common Crawl corpora for both directions and WMT 10^9 for English→French and the additional monolingual training data. The big noisy 10^9 and Crawl corpora were filtered using an SVM classifier [42]. In addition to the standard preprocessing, KIT used compound splitting [29] for the German text.

In both translation directions, KIT performed reordering using two models. KIT encoded different reorderings of the source sentences in a word lattice. For the English→French system, only short-range rules [43] were used to generate these lattices. For German→English, long-range rules [44] and tree-based reordering rules [45] were used as well. The part-of-speech (POS) tags needed for these rules were generated by the TreeTagger [46] and the parse trees by the Stanford Parser [47]. In addition, KIT scored the different reorderings of both language pairs using a lexicalized reordering model [48].

The phrase tables of the systems were trained using GIZA++ alignment for the English→French task and using a discriminative alignment [49] for the German→English task. KIT adapted the phrase table to the TED domain using the back off approach and by also adapting the candidate selection [50]. In addition to the phrase table probabilities, KIT modeled the translation process by a bilingual language model [51] and a discriminative word lexicon [52]. For the German→English task, a discriminative word lexicon with source and target context features was applied, while only the source context features were employed for the English→French task.

During decoding, KIT used several language models to adapt the system to the task and to better model the sentence structure by means of class-based *n*-grams. For the German→English task, KIT used one language model trained on all data, an in-domain language model trained only on the WIT^3 corpus and one language model trained on 5 M sentences selected using cross-entropy difference [23]. Furthermore, KIT used an RBM-based language model [53] trained on the WIT^3 corpus. Finally, KIT also used a class-based language model, trained on the WIT^3 corpus using the MKCLS [54] algorithm to cluster the words. For the English→French translation task, KIT linearly combined the language models trained on WIT^3, Europarl, News Commentary, 10^9, and Common Crawl by minimizing the perplexity on the development data. For the class-based language model, KIT utilized in-domain WIT^3 data with 4-grams and 50 clusters. In addition, a 9-gram POS-based language model derived from LIA POS tags [55] on all monolingual data was applied.

KIT optimized the log-linear combination of all these models on the provided development data using Minimum Error Rate Training [20].

5. Fondazione Bruno Kessler

The FBK component of the system combination corresponds to the "contrastive 1" system of the official FBK submission. The FBK system was built upon a standard phrase-based system using the Moses toolkit [32], 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 filled-up phrase translation model [56] and lexicalized reordering models (RMs), two French language models (LMs), as well as distortion, word, and phrase penalties. In order to focus it on TED specific domain and genre, and to reduce the size of the system, data selection by means of IRSTLM toolkit [57] was performed on the whole parallel English→French corpus, using the WIT^3 training data as in-domain data. Different amount of data are selected from each available corpora but the WIT^3 data, for a total of 66 M English running words. Two TMs and two RMs were trained on WIT^3 and selected data, separately, and combined using the fill-up (for TM) and back-off (for RM) techniques, using WIT^3 as primary component. The French side of WIT^3 and selected data were employed to estimate a mixture language model [58]. A second huge French LM was estimated on the monolingual French available data of about 2.4 G running words. Both LMs have order five and were smoothed by means of the interpolated Improved Kneser-Ney method [59]; the second LM was also pruned-out of singleton *n*-gram ($n > 2$). Tuning of the system was performed on dev2010 by optimizing BLEU using Minimum Error Rate Training [20]. It is worth noticing that the dev2010 and test2010 data were added to the training data in order to build the system actually employed in the translation of test2011, test2012, test2013.

6. 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.

The basic concept of RWTH's approach to machine translation system combination has been described by Matusov et al. [60]. This approach includes an enhanced alignment and reordering framework. Alignments between the system outputs are learned using METEOR [61]. 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.

7. Results

In this section, we present our experimental results on the two translation tasks, German→English and English→French.

7.1. German→English

RWTH Aachen University, the University of Edinburgh, and Karlsruhe Institute of Technology participated in the German→English translation task. The individual results as well as the system combination results are given in Table 1. RWTH's phrase-based translation (*scss*) is the best of the four single systems on test2010. The pairwise difference of the single system performance is up to 1.5 points in BLEU. In the end each system was needed to reach the performance of our final system combination submission. We optimized our system combination parameters on test2010. With the standard set of features, we got a gain of 1.5 BLEU on dev2010 and 1.2 BLEU on test2010 compared to the best single system. We tried different setups; also one which includes the large language model from RWTH's single systems as additional language model (+ *bigLM*). The translation quality in terms of BLEU improves by 0.2 on test2010 but degrades by 0.4 on dev2010. The TER scores were improved on both test sets, though. We decided to submit the system combination including *bigLM* as primary submission and the system combination without the large language model as secondary submission.

7.2. English→French

RWTH Aachen University, the University of Edinburgh, Karlsruhe Institute of Technology, and Fondazione Bruno Kessler participated in the English→French translation task. In Table 2 the results of the individual systems and our best system combination results are listed. The best individual system was provided by UEDIN. In this language pair the pairwise difference of the single systems was up to 1.5 points in BLEU. As in the German→English translation task, we tried to optimize our parameters on test2010. We got a large improvement on test2010 of 2.1 points in BLEU, but got only a slight improvement of 0.3 BLEU on dev2010. After changing the optimization set to dev2010, we got comparable improvements on both test sets. On dev2010 we got an improvement of 1.4 points in BLEU and on test2010 an improvement of 0.8 points in BLEU. On both test sets the performance in TER was similar or even better compared to the system combination optimized on test2010. We decided to submit the system combination optimized on dev2010 as primary submission.

Table 1: Results for the German→English translation task. Bold font indicates system combination results that are significantly better than the best single system ($p < 0.05$).

system	dev2010		test2010	
	BLEU	TER	BLEU	TER
RWTH scss	33.3	47.0	31.4	49.3
KIT	33.6	46.5	31.1	49.5
RWTH hiero	33.0	46.8	30.7	49.5
UEDIN	32.1	47.3	29.9	49.6
sc	**34.8**	**44.9**	**32.6**	**47.4**
sc + bigLM	**34.4**	**44.4**	**32.8**	**46.5**

Table 2: Results for the English→French translation task. Bold font indicates system combination results that are significantly 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 $p < 0.1$.

system	dev2010		test2010	
	BLEU	TER	BLEU	TER
UEDIN	29.4	55.4	33.2	49.8
RWTH scss	28.8	55.4	32.8	49.2
KIT	28.8	55.7	32.6	49.3
FBK	27.5	57.0	32.1	50.0
RWTH hiero	28.0	56.3	31.7	49.9
sc opt dev10	*30.8*	**53.8**	*34.0*	**48.1**
sc opt test10	29.7	55.2	**35.3**	**48.2**

8. Error Analysis

We carried out a restricted manual error analysis to compare the outputs of each single system to the final system combination output for some example sentences. In Figure 1 and Figure 2 the TER scores of all translations of two selected sentences from the German→English translation direction are given. In both sentences system combination outperforms each single system.

KIT (TER Score: 60.00 (9.0/ 15.0))	
hyp	except for your contribution , whatever it may be .
shifted hyp	——— – —- except for your —— contribution it , whatever —- may be .
edited hyp	——— – —- except for your —— contribution it , whatever —- may be .
ref	continue to show up for your piece of it , whatever that might be .
RWTH hiero (TER Score: 66.67 (10.0/ 15.0))	
hyp	is still there for the post , whatever it may be .
shifted hyp	——— is still there for —- the post it , whatever —- may be .
edited hyp	——— is still there for —- the post it , whatever —- may be .
ref	continue to show up for your piece of it , whatever that might be .
RWTH scss (TER Score: 66.67 (10.0/ 15.0))	
hyp	for your contribution is still there , whatever it may be .
shifted hyp	——— – —- – for your contribution is still there , whatever it may be .
edited hyp	——— – —- – for your contribution is still there , whatever it may be .
ref	continue to show up for your ———— piece of it , whatever that might be .
UEDIN (TER Score: 66.67 (10.0/ 15.0))	
hyp	continue to be there for your contribution , which may be his time .
shifted hyp	continue to —- there for your contribution which may , be his time be .
edited hyp	continue to —- there for your contribution which may , be his time be .
ref	continue to show up for your piece of it , whatever that might be .
system combination (TER Score: 46.67 (7.0/ 15.0))	
hyp	continue to be there for your contribution , whatever it may be .
shifted hyp	continue to be there for your —— contribution it , whatever —- may be .
edited hyp	continue to be there for your —— contribution it , whatever —- may be .
ref	continue to show up for your piece of it , whatever that might be .

Figure 1: Error analysis for sentence 715 (dev2010) in the German→English translation task.

KIT (TER Score: 52.63 (10.0/ 19.0))	
hyp	they can only be in conjunction with a number of other chemicals taken the mao advised .
shifted hyp	they can only be —— ——— – taken in conjunction with a other number of chemicals the mao advised .
edited hyp	they can only be —— ——— – taken in conjunction with a other number of chemicals the mao advised .
ref	they can only be taken orally if taken in conjunction with some other chemical that denatures the mao ———- .
RWTH hiero (TER Score: 47.37 (9.0/ 19.0))	
hyp	they can only be taken in combination with other chemicals , who turned the mao .
shifted hyp	they can only be —— ——— – taken in combination with chemicals other , who turned the mao .
edited hyp	they can only be —— ——— – taken in combination with chemicals other , who turned the mao .
ref	they can only be taken orally if taken in conjunction with some other chemical that denatures the mao .
RWTH scss (TER Score: 47.37 (9.0/ 19.0))	
hyp	they can only be consumed in connection with some other chemicals , which turned the mao .
shifted hyp	they can only be —— ——— – consumed in connection with some other chemicals , which turned the mao .
edited hyp	they can only be —— ——— – consumed in connection with some other chemicals , which turned the mao .
ref	they can only be taken orally if taken in conjunction with some other ———— chemical that denatures the mao .
UEDIN (TER Score: 36.84 (7.0/ 19.0))	
hyp	they can be used only in conjunction with other chemicals taken orally that denaturieren the mao .
shifted hyp	they can only be taken orally – used in conjunction with —- other chemicals that denaturieren the mao .
edited hyp	they can only be taken orally – used in conjunction with —- other chemicals that denaturieren the mao .
ref	they can only be taken orally if taken in conjunction with some other chemical that denatures the mao .
system combination (TER Score: 26.32 (5.0/ 19.0))	
hyp	they can only be taken in conjunction with other chemicals taken orally that turned the mao .
shifted hyp	they can only be taken orally – taken in conjunction with —- other chemicals that turned the mao .
edited hyp	they can only be taken orally – taken in conjunction with —- other chemicals that turned the mao .
ref	they can only be taken orally if taken in conjunction with some other chemical that denatures the mao .

Figure 2: Error analysis for sentence 90 (dev2010) in the German→English translation task.

Words marked with red are substitutions, blue are insertions, green are deletions and yellow are shifts. In Figure 1 the final system combination translation is build out of the beginning part of UEDIN and the end part of all other single systems. Combined, this new translation improves over all single systems in terms of TER. In Figure 2 the system combination output is basically a fixed version of the UEDIN translation. This results in a better TER score which needs two less edits.

9. Conclusion

For our participation in the MT track of the IWSLT 2013 evaluation campaign, four partners from the EU-BRIDGE project (RWTH Aachen University, University of Edinburgh, Karlsruhe Institute of Technology, Fondazione Bruno Kessler) provided a joint submission. Our combined EU-BRIDGE system setup for text translation of talks is part of our efforts within the project to deliver high-quality machine translation of spoken language.

By joining the outputs of the partners' different individual machine translation engines via a system combination framework we have been able to achieve significantly better translation performance (up to +1.4 BLEU and -2.8 TER). While each of the individual engines provides performance that is state-of-the-art for single systems, our results suggest that system combination techniques are still a fertile approach to benefit from diversity in collaborative efforts and thus progress towards even better quality.

In future research we intend to both improve single systems and to investigate novel methods and models in machine translation system combination for large-scale and real-world settings.

10. Acknowledgements

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

11. References

[1] International Workshop on Spoken Language Translation 2013, http://www.iwslt2013.org.

[2] TED Talks, http://www.ted.com/talks.

[3] M. Federico, L. Bentivogli, M. Paul, and S. Stueker, "Overview of the IWSLT 2011 Evaluation Campaign," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, USA, Dec. 2011.

[4] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 Evaluation Campaign," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, December 2012.

[5] M. Cettolo, C. Girardi, and M. Federico, "WIT^3: Web Inventory of Transcribed and Translated Talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[6] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proc. of the MT Summit X*, Phuket, Thailand, Sept. 2005.

[7] A. Eisele and Y. Chen, "MultiUN: A Multilingual Corpus from United Nation Documents," in *Proceedings of the Seventh conference on International Language Resources and Evaluation*, May 2010, pp. 2868–2872.

[8] Linguistic Data Consortium (LDC), http://www.ldc.upenn.edu.

[9] Shared Translation Task of the ACL 2013 Eighth Workshop on Statistical Machine Translation, http://www.statmt.org/wmt13/translation-task.html.

[10] European Commission Community Research and Development Infomation Service (CORDIS), "Seventh Framework Programme (FP7)," http://cordis.europa.eu/fp7/.

[11] E. Matusov, N. Ueffing, and H. Ney, "Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment," in *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2006, pp. 33–40.

[12] M. Freitag, S. Peitz, M. Huck, H. Ney, T. Herrmann, J. Niehues, A. Waibel, A. Allauzen, G. Adda, B. Buschbeck, J. M. Crego, and J. Senellart, "Joint WMT 2012 Submission of the QUAERO Project," in *NAACL 2012 Seventh Workshop on Statistical Machine Translation*, Montréal, Canada, June 2012, pp. 322–329.

[13] S. Peitz, S. Mansour, M. Huck, M. Freitag, H. Ney, E. Cho, T. Herrmann, M. Mediani, J. Niehues, A. Waibel, A. Allauzen, Q. K. Do, B. Buschbeck, and T. Wandmacher, "Joint WMT 2013 Submission of the QUAERO Project," in *ACL 2013 Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, Aug. 2013.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA, July 2002, pp. 311–318.

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

[16] 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.

[17] Vilar, David and Stein, Daniel and Huck, Matthias and Ney, Hermann, "Jane: an advanced freely available hierarchical machine translation toolkit," *Machine Translation*, vol. 26, no. 3, pp. 197–216, Sept. 2012.

[18] M. Huck, J.-T. Peter, M. Freitag, S. Peitz, and H. Ney, "Hierarchical Phrase-Based Translation with Jane 2," *The Prague Bulletin of Mathematical Linguistics (PBML)*, vol. 98, pp. 37–50, Oct. 2012.

[19] 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.

[20] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, July 2003, pp. 160–167.

[21] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, vol. 2, Denver, CO, USA, Sept. 2002, pp. 901–904.

[22] C. Dyer, V. Chahuneau, and N. A. Smith, "A Simple, Fast, and Effective Reparameterization of IBM Model 2," in *Proceedings of NAACL-HLT*, Atlanta, GA, USA, June 2013, pp. 644–648.

[23] R. Moore and W. Lewis, "Intelligent Selection of Language Model Training Data," in *ACL (Short Papers)*, Uppsala, Sweden, July 2010, pp. 220–224.

[24] 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.

[25] M. Galley and C. D. Manning, "A Simple and Effective Hierarchical Phrase Reordering Model," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08, Honolulu, HI, USA, 2008, pp. 848–856.

[26] A. Mauser, S. Hasan, and H. Ney, "Extending statistical machine translation with discriminative and trigger-based lexicon models," in *Conference on Empirical Methods in Natural Language Processing*, Singapore, Aug. 2009, pp. 210–217.

[27] J. Wuebker, S. Peitz, F. Rietig, and H. Ney, "Improving Statistical Machine Translation with Word Class Models," in *Conference on Empirical Methods in Natural Language Processing*, Seattle, WA, USA, Oct. 2013, pp. 1377–1381.

[28] H. Schwenk, A. Rousseau, and M. Attik, "Large, Pruned or Continuous Space Language Models on a GPU for Statistical Machine Translation," in *NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, Montréal, Canada, June 2012, pp. 11–19.

[29] P. Koehn and K. Knight, "Empirical Methods for Compound Splitting," in *Proc. 10th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL)*, Budapest, Hungary, Apr. 2003, pp. 347–354.

[30] 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.

[31] X. He and L. Deng, "Maximum Expected BLEU Training of Phrase and Lexicon Translation Models," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, Jeju, Republic of Korea, Jul 2012, pp. 292–301.

[32] 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 *ACL 2007 Demonstrations*, Prague, Czech Republic, 2007.

[33] N. Durrani, B. Haddow, K. Heafield, and P. Koehn, "Edinburgh's Machine Translation Systems for European Language Pairs," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August 2013.

[34] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, UK, July 2011, pp. 187–197.

[35] N. Durrani, H. Schmid, and A. Fraser, "A Joint Sequence Translation Model with Integrated Reordering," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, OR, USA, June 2011, pp. 1045–1054.

[36] 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.

[37] S. Kumar and W. Byrne, "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)*, Boston, MA, USA, May 2004, pp. 169–176.

[38] L. Huang and D. Chiang, "Forest Rescoring: Faster Decoding with Integrated Language Models," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, June 2007, pp. 144–151.

[39] M. Junczys-Dowmunt, "Phrasal Rank-Encoding: Exploiting Phrase Redundancy and Translational Relations for Phrase Table Compression," *The Prague Bulletin of Mathematical Linguistics*, vol. 98, pp. 63–74, 2012.

[40] C. Cherry and G. Foster, "Batch Tuning Strategies for Statistical Machine Translation," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, June 2012, pp. 427–436.

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

[42] M. Mediani, E. Cho, J. Niehues, T. Herrmann, and A. Waibel, "The KIT English-French Translation systems for IWSLT 2011," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, 2011.

[43] 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.

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

[45] 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.

[46] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," in *International Conference on New Methods in Language Processing*, Manchester, United Kingdom, 1994.

[47] A. N. Rafferty and C. D. Manning, "Parsing three german treebanks: lexicalized and unlexicalized baselines," in *Proceedings of the Workshop on Parsing German*, 2008.

[48] P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot, "Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Pittsburgh, PA, USA, 2005.

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

[50] J. Niehues and A. Waibel, "Detailed Analysis of different Strategies for Phrase Table Adaptation in SMT," in *Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, San Diego, CA, USA, Oct. / Nov. 2012.

[51] J. Niehues, T. Herrmann, S. Vogel, and A. Waibel, "Wider Context by Using Bilingual Language Models in Machine Translation," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, 2011.

[52] 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, Aug. 2013, pp. 512–520.

[53] ——, "Continuous Space Language Models using Restricted Boltzmann Machines," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, Dec. 2012.

[54] F. J. Och, "An Efficient Method for Determining Bilingual Word Classes." in *EACL'99*, 1999.

[55] F. Béchet, "LIA PHON: Un Systeme Complet de Phonétisation de Textes," *Traitement automatique des langues*, vol. 42, no. 1, pp. 47–67, 2001.

[56] 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.

[57] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an open source toolkit for handling large scale language models," in *Interspeech*, 2008, pp. 1618–1621.

[58] M. Federico and R. De Mori, "Language modelling," *Spoken Dialogues with Computers*, pp. 199–230, 1998.

[59] F. James, "Modified Kneser-Ney Smoothing of n-gram Models," RIACS, Tech. Rep. 00.07, Oct. 2000.

[60] E. Matusov, G. Leusch, R. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y.-S. Lee, J. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney, "System Combination for Machine Translation of Spoken and Written Language," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 7, pp. 1222–1237, 2008.

[61] 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.