The RWTH Aachen Machine Translation Systems for IWSLT 2013

Joern Wuebker, Stephan Peitz, Tamer Alkhouli, Jan-Thorsten Peter
Minwei Feng, Markus Freitag and Hermann Ney

Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
Aachen, Germany
<surname>@cs.rwth-aachen.de

Abstract

This work describes the statistical machine translation (SMT) systems of RWTH Aachen University developed for the evaluation campaign *International Workshop on Spoken Language Translation* (IWSLT) 2013. We participated in the English→French, English↔German, Arabic→English, Chinese→English and Slovenian↔English MT tracks and the English→French and English→German SLT tracks. We apply phrase-based and hierarchical SMT decoders, which are augmented by state-of-the-art extensions. The novel techniques we experimentally evaluate include discriminative phrase training, a continuous space language model, a hierarchical reordering model, a word class language model, domain adaptation via data selection and system combination of standard and reverse order models. By application of these methods we can show considerable improvements over the respective baseline systems.

1. Introduction

We describe the statistical machine translation (SMT) systems developed by RWTH Aachen University for the evaluation campaign of IWSLT 2013. We participated in the machine translation (MT) track for the language pairs English→French, English↔German, Arabic→English, Chinese→English and Slovenian↔English and the spoken language translation (SLT) tracks for the language pairs English→French and English→German. We apply state-of-the-art phrase-based and hierarchical machine translation systems as well as an in-house system combination framework. To improve the baselines, we evaluated several different methods in terms of translation performance. These include a discriminative phrase training technique, continuous space language models, a hierarchical reordering model for the phrasal decoder, word class (cluster) language models, domain adaptation via data selection, application of two separate translation models or phrase table interpolation, word class translation and reordering models, optimization with PRO and a discriminative word lexicon. Further, on the small scale Slovenian↔English tasks we compare the performance of the two word alignment toolkits GIZA++ and *fast_align*. For the spoken language translation task, the ASR output is enriched with punctuation and casing. The enrichment is performed by a hierarchical phrase-based translation system.

This paper is organized as follows. In Section 2 we describe our translation software and baseline setups. Sections 2.3 and 2.4 introduce the novel discriminative phrase training technique and the continuous space language model, whose application shows improvements on several tasks. Our experiments for each track are summarized in Section 3 and we conclude with Section 4.

2. SMT Systems

For the IWSLT 2013 evaluation campaign, RWTH utilized state-of-the-art phrase-based and hierarchical translation systems as well as our in-house system combination framework. GIZA++ [1] or *fast_align* [2] are employed to train word alignments. All language models are created with the SRILM toolkit [3] and are standard 4-gram LMs with interpolated modified Kneser-Ney smoothing. We evaluate in case-insensitive fashion, using the BLEU [4] and TER [5] measures.

2.1. Phrase-based Systems

As phrase-based SMT systems, in this work we used both an in-house implementation of the state-of-the-art MT decoder (PBT) described in [6] and the implementation of the decoder based on [7] (SCSS) which is part of RWTH's open-source SMT toolkit Jane 2.1 [1]. We use the standard set of models with phrase translation probabilities and lexical smoothing in both directions, word and phrase penalty, distance-based reordering model, an n-gram target language model and three binary count features. The parameter weights are optimized with MERT [8], PRO [9] (SCSS) or the downhill simplex algorithm [10] (PBT).

Additional state-of-the-art models that are applied successfully in the IWSLT 2013 evaluation are a hierarchi-

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

cal reordering model (HRM) [11], a high-order word class language model (wcLM) [12], word class based translation and reordering models (wcTM) [12], a discriminative phrase training scheme (cf. Section 2.3) and rescoring with a neural network language model (cf. Section 2.4).

2.2. Hierarchical Phrase-based System

For our hierarchical setups, we employed the open source translation toolkit Jane [13], which has been developed at RWTH and is freely available for non-commercial use. In hierarchical phrase-based translation [14], a weighted synchronous context-free grammar is induced from parallel text. In addition to contiguous *lexical* phrases, *hierarchical* phrases with up to two gaps are extracted. The search is carried out with a parsing-based procedure. The standard models integrated into our Jane systems are: phrase translation probabilities and lexical smoothing probabilities in both translation directions, word and phrase penalty, binary features marking hierarchical phrases, glue rule, and rules with non-terminals at the boundaries, four binary count features, phrase length ratios and an n-gram language model. We utilize the cube pruning algorithm [15] for decoding and optimize the model weights with standard MERT [8] on 100-best lists.

2.3. Discriminative Phrase Training

The state of the art for creating the phrase tables of standard SMT systems is still a heuristic extraction from word alignments and probability estimation as relative frequencies. In several systems for the IWSLT 2013 shared task, we applied a more sophisticated discriminative phrase training method. Similar to [16], 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. In the experiments reported in this paper, we perform discriminative training on the TED portion of the training data in all cases. To that end, we decode the training data to generate 100-best lists. A leave-one-out heuristic [17] is applied to make better use of the training data. Using these n-best lists, we iteratively perform updates on the phrasal translation scores of the phrase table. After each iteration, we perform MERT, evaluate on the development set and finally select the iteration which performs best.

2.4. Neural Network Language Model

We train neural networks as language models using the theano numerical computation library[18]. The neural network structure is largely similar to the continuous space language model (*CSLM*) [19]. Our input layer includes a short list of the most common word and word factors like the word beginning or ending. To reduce the computation cost of the network we employ a clustered output layer [20, 21]. The Neural Network Language Model is used as a final step in our translation pipeline, by rescoring on 200-best lists for the English→French and English→German tasks.

Table 1: Results for the English→French MT task.

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
SCSS allData	28.3	55.7	31.9	49.8
+HRM	28.7	55.3	32.5	49.2
+2TM	29.2	54.7	32.7	48.9
+GW	29.5	54.6	32.9	48.9
+DWL	29.8	54.3	33.2	48.5
+wcLM	29.7	54.2	33.5	48.3
+CSLM	30.0	53.8	33.7	48.0

3. Experimental Evaluation

3.1. English→French

For the English→French task, the word alignment was trained with GIZA++ and we applied the phrase-based decoder implemented in Jane. We used all available parallel data for training the translation model. The baseline French LM is trained on the target side of all available bilingual data plus $\frac{1}{2}$ of the Shuffled News corpus. The monolingual data selection is based on cross-entropy difference as described in [22]. The experimental results are given in Table 1. Different from last year [23], we did not employ system combination in this task, achieving similar results with a single decoder. The baseline system is improved by the hierarchical reordering model (HRM, +0.6% BLEU), adding a second translation model to the decoder (2TM, +0.2% BLEU), which was trained on the TED portion of the data, using $\frac{1}{4}$ of the French Gigaword Second Edition corpus as additional language model training data (GW, +0.2% BLEU), and smoothing the translation model with a discriminative word lexicon [24] trained on the in-domain data (+0.3% BLEU). For the final submission, we applied two additional language models: the 7-gram word class language model (wcLM, 0.3% BLEU) and the neural language model (CSLM, 0.2% BLEU).

3.2. German↔English

Similar to English→French, for the German↔English tasks, we used GIZA++ for the word alignments and applied the phrase-based decoder from the Jane toolkit.

For the German→English translation direction, in a preprocessing step the German source is decompounded [25] and part-of-speech-based long-range verb reordering rules [26] are applied. The English LM is trained the target side of all available bilingual data plus a selection [22] of $\frac{1}{2}$ from the Shuffled News corpus and $\frac{1}{4}$ from the English Gigaword v3 corpus, resuling in a total of 1.7 billion running words. The experimental results for the German→English task are given in Table 2. In opposition to our findings from last

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

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
SCSS TED	31.5	47.6	30.0	49.2
SCSS allData	32.8	46.4	30.3	48.9
+HRM	33.0	46.1	30.4	48.9
+wcLM	33.5	45.8	30.9	48.4
+discr.	33.9	45.0	31.4	47.5
+2TM	34.2	45.2	32.3	47.4

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

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
SCSS TED	22.0	56.7	21.9	57.3
SCSS allData	22.7	56.1	22.3	57.2
+HRM	23.3	55.5	22.6	57.7
+wcLM	24.2	54.5	23.6	55.9
+discr.	24.6	54.1	24.3	55.4
+CSLM	24.7	53.7	24.9	54.7

year [23], using all available data now performs better than solely training on the in-domain TED portion. This can be attributed to the large, newly available Common Crawl corpus. The baseline system is improved by the hierarchical reordering model (HRM, +0.1% BLEU), the 7-gram word class language model (wcLM, 0.5% BLEU) and discriminative phrase training (discr., +0.5% BLEU). Finally, we applied domain adaptation by adding a second translation model to the decoder (2TM), which was trained on the TED portion of the data. This second translation model was also trained with discriminative phrase training and gave an additional improvement of 0.9% BLEU.

The English→German system is very similar to the one for the opposite translation direction. The language model was trained on the target side of all bilingual data plus $\frac{1}{2}$ of the Shuffled News corpus selected with [22]. The LM training data contains a total of 564 million running words. The results in Table 3 show that using all available training data outperforms only training on the in-domain TED portion. The system is augmented with the hierarchical reordering model (HRM, +0.3% BLEU), a word class language model (wcLM, 1.0% BLEU) and disriminative phrase training (discr., +0.5% BLEU). Especially the wcLM has a strong impact on translation performance. Different from the opposite direction, adding a second translation model did not improve results. However, we were able to reach a final improvement of 0.6% BLEU by rescoring a 200-best list with a neural language model (CSLM).

3.3. Arabic→English

The Arabic→English system uses a language model based on the full in-domain TED and out-of-domain UN and News Commentary v8 data. We also filtered and included the English Gigaword, giga-fren.en, Europarl v7, Common Crawl and Shuffled News corpora using the cross-entropy criterion. A 4-gram LM is trained for each of the sets using modified Kneser-Ney discounting with interpolation. The final LM is the weighted mixture of all individual LMs, with the weights tuned to achieve the lowest perplexity on dev2010. We also trained another mixture of LMs keeping singleton *n*-grams, which we will refer to as sngLM.

A single system employing MADA v3.1 D3 resulted in only 0.3% worse BLEU and TER on the tst2011 dataset of IWSLT2012, compared to a system combination where single systems of various segmentation techniques were combined, as described in [23]. Therefore, we stuck to a single system using MADA v3.1 D3 for segmentation. The translation model is trained using the TED and UN bilingual corpora, and the standard features were used in addition to HRM. Two phrase tables were built, one based on the TED dataset and the other on the TED+UN data. We interpolated the two linearly with the weights 0.9 and 0.1 respectively given to the TED and the full phrase tables. Table 4 shows the results. The HRM features bring an improvement of 1.1% BLEU and 0.2% TER to a TED-only translation model. Adding the UN data hurts performance by 1.1% BLEU and 0.7% TER. On the other hand, interpolation leads to an improvement of 0.8% in TER and 0.1% BLEU. When replacing the LM with sngLM an improvement is only observed on the development set, but not the test set, which could not be remedied by relaxing the pruning parameters. All sngLM experiments used a 200-best list, compared to a 100-best list used with the smaller LM.

We also experimented with bilingual filtering of the UN data used to train the phrase table, where scoring was performed using bilingual LM cross-entropy scores (x-entropy) [27]. Another experiment used the combination of cross-entropy and IBM-1 scores (x-entropy+IBM-1) [28]. We used the best 400k UN sentences together with the TED data to train a phrase table, which is then interpolated with a TED-only phrase table as described above. x-entropy+IBM-1 is better by 0.8% TER than mere cross-entropy filtering, and it performs similar to the non-filtered system, despite the fact that we select only $\frac{1}{16}$ of the UN data.

3.4. Chinese→English

For the Chinese-English task, RWTH utilized system combination as described in [29]. We used both the phrase-based decoder and the hierarchical phrase-based decoder to perform a bi-directional translation, which means the system performs standard direction decoding (left-to-right) and reverse direction decoding (right-to-left). To build the reverse direction system, we used exactly the same data as the stan-

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

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
SCSS TED	27.4	52.0	25.7	55.1
+HRM	27.9	51.9	26.8	54.9
+UN	28.4	51.9	25.7	55.6
+UN interpolated	28.3	51.1	26.9	54.1
+sngLM	28.8	50.7	26.8	54.1
+x-entropy	28.6	51.8	26.7	55.0
+x-entropy+IBM-1	28.8	51.0	27.0	54.2

Table 5: Chinese-English results on the dev test set for different segmentations. The primary submission is a system combination of all the listed systems.

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
PBT-2012-standard	11.5	80.7	13.0	76.4
PBT-2012-reverse	11.7	80.9	13.6	75.5
HPBT-2012-standard	12.3	79.8	14.2	74.6
HPBT-2012-reverse	12.8	79.4	14.6	74.1
HPBT-2013-standard	12.4	79.5	14.5	74.1
HPBT-2013-reverse	12.6	79.4	14.4	74.3
system combination	13.5	78.5	15.1	73.6

dard direction system and simply reversed the word order of the bilingual corpora. For the system combination we selected four systems we had trained for last year's IWSLT evaluation and set up two additional hierarchical systems with slightly different preprocessing. Note that all translation model are trained on the in-domain data only. By performing system combination we gain an improvement of +0.5% BLEU over the best single system. Results are given in Table 5.

3.5. Slovenian↔English

The bilingual training data available for the Slovenian↔English tasks is limited to 14K sentence pairs from the TED lecture domain. Further, only one development set was provided. In order to be able to do blind evaluation, we split it into two parts. The first 644 lines are defined as `dev1` and are used for MERT/PRO. The remaining 500 lines are used as blind test set and will be referred to as `dev2`. For the Slovenian↔English tasks, we apply our phrase-based decoder and experimented with two different word alignments for training, one generated with GIZA++, based on the IBM model 4, and one created with *fast_align*, which uses a reparameterization of IBM model 2. Interestingly, the simpler and more efficient *fast_align* tool outperforms GIZA++ in both cases.

Table 6: Results for the Slovenian→English MT task. All systems are augmented with the hierarchical reordering model.

system	dev1		dev2	
	BLEU	TER	BLEU	TER
SCSS GIZA++	17.6	65.7	15.9	67.6
SCSS *fast_align*	18.0	64.8	16.3	66.1
+wcLM	18.2	62.9	16.5	64.6
+wcTM +PRO	18.6	63.0	16.5	64.3
+discr.	18.8	62.6	16.9	63.9

Table 7: Results for the English→Slovenian MT task. All systems are augmented with the hierarchical reordering model.

system	dev1		dev2	
	BLEU	TER	BLEU	TER
SCSS GIZA++	11.3	70.5	9.6	71.4
SCSS *fast_align*	11.4	70.3	10.5	69.6
+wcLM	12.0	69.8	10.1	69.9
+wcTM	11.9	70.3	10.4	69.9
+discr.	11.9	70.2	10.7	69.7

The Slovenian→English MT system uses the same language model as described in Section 3.2 for the German→English task. Results are shown in Table 6. The baseline, which already contains the hierarchical reordering model, is augmented with a word class LM (wcLM, +0.2% BLEU) and the word class translation and reordering model (wcTM). When we add the latter, we switch from MERT to PRO, which we found to lead to more stable results in this case. Finally, we employ discriminative phrase training (discrim., +0.4% BLEU) to build the submission system.

To train the Slovenian language model only the target side of the bilingual data was provided. We found that selecting a submission system on this task was very difficult, as when comparing two setups, their behaviour was often reversed between `dev1` and `dev2`. We decided to apply the same extensions to the baseline as for the opposite translation direction. The baseline, which already contains the hierarchical reordering model, is augmented with the word class LM and the word class based translation and reordering models. Here, we continue using MERT. For the final submission, we also applied discriminative phrase training. Results are shown in Table 7.

3.6. Spoken Language Translation (SLT)

RWTH participated in the English→French and English→German SLT task. In both tracks, we reintroduced punctuation and case information following [30],

Table 8: Results for the English→French SLT task.

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
	23.0	62.7	26.0	56.0
re-optimized	23.4	62.5	26.3	56.2

which we denote as *enriched*. Further, we added a phrase feature, that fires if a phrase introduces a punctuation mark on the target side. The SMT system, that is employed in the enrichment process by translating from pure ASR output to the enriched version, we use a hierarchical phrase-based system with a maximum of one nonterminal symbol per rule The model weights are tuned with standard MERT on 100-best lists. As optimization criterion we use WER.

For English→French, we re-optimized on the enriched ASR development dev using **SCSS allData** +HRM +GW +2TM. Results are reported in Table 8.

For English→German, the enriched evaluation set was translated using the **SCSS allData** +HRM +wcLM +discr. system. Here, the translation system was kept completely unchanged from the MT task, including the log-linear feature weights.

4. Conclusion

RWTH participated in seven MT tracks and two SLT tracks of the IWSLT 2013 evaluation campaign. The baseline systems utilize our state-of-the-art translation decoders and we were able to improve them by applying novel models or techniques. The most notable improvements are achieved by a hierarchical reordering model (+1.1 BLEU on Ar-En), a word class language model (+1.0 BLEU on En-De), discriminative phrase training (+0.7 BLEU on En-De), a continuous space language model (+0.6 BLEU on En-De) and system combination of standard and reverse order models (+0.5 BLEU on Zh-En). For the SLT track, the ASR output was enriched with punctuation and casing information by a hierarchical translation system.

5. Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 287658 and n° 287755. This work was also partly achieved as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

6. References

[1] 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.

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

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

[4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, July 2002, pp. 311–318.

[5] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, August 2006, pp. 223–231.

[6] R. Zens and H. Ney, "Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation," in *International Workshop on Spoken Language Translation*, Honolulu, Hawaii, Oct. 2008, pp. 195–205.

[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 *International Conference on Computational Linguistics*, Mumbai, India, Dec. 2012, to appear.

[8] 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.

[9] M. Hopkins and J. May, "Tuning as ranking," in *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, July 2011, pp. 1352–1362.

[10] J. A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *The Computer Journal*, vol. 7, pp. 308–313, 1965.

[11] 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. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 848–856. [Online]. Available: http://dl.acm.org/citation.cfm?id=1613715.1613824

[12] 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, USA, Oct. 2013, pp. 1377–1381.

[13] 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.

[14] D. Chiang, "Hierarchical Phrase-Based Translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

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

[16] 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.

[17] J. Wuebker, A. Mauser, and H. Ney, "Training phrase translation models with leaving-one-out," in *Proceedings of the 48th Annual Meeting of the Assoc. for Computational Linguistics*, Uppsala, Sweden, July 2010, pp. 475–484.

[18] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010, oral Presentation.

[19] 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.

[20] J. Goodman, "Classes for fast maximum entropy training," *CoRR*, vol. cs.CL/0108006, 2001.

[21] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model," in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, R. G. Cowell and Z. Ghahramani, Eds. Society for Artificial Intelligence and Statistics, 2005, pp. 246–252. [Online]. Available: http://www.iro.umontreal.ca/~lisa/pointeurs/hierarchical-nnlm-aistats05.pdf

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

[23] S. Peitz, S. Mansour, M. Freitag, M. Feng, M. Huck, J. Wuebker, M. Nuhn, M. Nußbaum-Thom, and H. Ney, "The rwth aachen speech recognition and machine translation system for iwslt 2012," in *International Workshop on Spoken Language Translation*, Hong Kong, Dec. 2012, pp. 69–76. [Online]. Available: http://hltc.cs.ust.hk/iwslt/proceedings/paper_45.pdf

[24] 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.

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

[26] 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.

[27] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., July 2011, pp. 355–362.

[28] S. Mansour, J. Wuebker, and H. Ney, "Combining Translation and Language Model Scoring for Domain-Specific Data Filtering," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, Dec. 2011.

[29] M. Freitag, M. Feng, M. Huck, S. Peitz, and H. Ney, "Reverse word order models," in *Machine Translation Summit*, Nice, France, Sept. 2013.

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