The RWTH Aachen Machine Translation Systems for IWSLT 2014

Joern Wuebker, Stephan Peitz, Andreas Guta 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) 2014. We participated in both the MT and SLT tracks for the English→French and German→English language pairs. We apply the identical training pipeline and models on both language pairs. Our state-of-the-art phrase-based baseline systems are augmented with maximum expected BLEU training for phrasal, lexical and reordering models. Further, we apply rescoring with novel recurrent neural language and translation models. The same systems are used for the SLT track, where we additionally perform punctuation prediction on the automatic transcriptions employing hierarchical phrase-based translation. We are able to improve RWTH's 2013 evaluation systems by 1.7-1.8% BLEU absolute.

1. Introduction

We describe the statistical machine translation (SMT) systems developed by RWTH Aachen University for the evaluation campaign of IWSLT 2014. We participated in the machine translation (MT) track and the spoken language translation (SLT) track for the language pairs English→French as well as German→English. We apply the identical training pipeline and models on both language pairs using a state-of-the-art phrase-based system. The pipeline include a hierarchical reordering model, word class (cluster) language models, discriminative phrase training and rescoring with novel recurrent neural language and translation models. 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 2.4 and 2.5 provide further details about our discriminative phrase training and the recurrent neural network models. Our experiments for each track are summarized in Section 3 and we conclude with Section 4.

2. SMT Systems

For the IWSLT 2014 evaluation campaign, RWTH utilized state-of-the-art phrase-based and hierarchical translation systems. GIZA++ [1] is employed to train word alignments. We evaluate in case-insensitive fashion, using the BLEU [2] and TER [3] measures.

2.1. Phrase-based Systems

Our phrase based decoder is the implementation of the *source cardinality synchronous search* (SCSS) procedure described in [4] in RWTH's open-source SMT toolkit Jane 2.3 [1] [5]. 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, n-gram target language models and and enhanced low frequency feature [6]. The parameter weights are optimized with MERT [7] towards the BLEU metric. Additionally, we make use of a hierarchical reordering model (HRM) [8], a high-order word class language model (wcLM) [9], a discriminative phrase training scheme (cf. Section 2.4) and rescoring with recurrent neural network language and translation models (cf. Section 2.5).

2.2. Hierarchical Phrase-based System

For our hierarchical setups, we employed the open source translation toolkit Jane [10], which has been developed at RWTH and is freely available for non-commercial use. In hierarchical phrase-based translation [11], 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, extended low frequency feature and an n-gram language model. We utilize the cube

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

pruning algorithm [12] for decoding.

2.3. Backoff language models

Each translation system uses three backoff language models that are estimated with the KenLM toolkit [13]: 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 use interpolated Kneser-Ney smoothing. For the general domain LM, we first select $\frac{1}{2}$ 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 [14]. The selection is then concatenated with all available remaining monolingual data and used to build and unpruned language model. The in-domain language models are estimated on the TED data only. For the word class LM, we train 200 classes on the target side of the bilingual training data using an in-house tool similar to `mkcls`. With these class definitions, we apply the technique shown in [9] to compute the wcLM on the same data as the general-domain LM.

2.4. Maximum Expected BLEU Training

Discriminative training is a powerful method to learn a large number of features with respect to a given error metric. In this work we learn three types of features under a maximum expected BLEU objective [15]. We perform discriminative training 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 [15], we generate 100-best lists on the training data which are used as training samples for a gradient based update method. A leave-one-out heuristic [16] is applied to circumvent over-fitting. Here, we follow an approach similar to [17], where each feature type is first discriminatively trained, then condensed into a single feature for the log-linear model combination and finally optimized with MERT. In a first pass, we simultaneously train phrase pair features and phrase-internal word pair features, adding two models to the log-linear combination. Afterwards we perform a second pass focusing on reordering, with the identical feature set as the HRM, resulting in an additional six models for log-linear combination: Three orientation classes (monotone, swap and discontinuous) in both directions. As the training procedure is iterative, we select the best iteration after performing MERT.

2.5. Recurrent Neural Network Models

All systems apply rescoring on 1000-best lists using recurrent language and translation models. The recurrency is handled with the long short-term memory (LSTM) architecture [18] and we use a class-factored output layer for increased efficiency as described in [19]. 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), we apply the deep bidirectional word-based translation model (RNN-BTM) described in [20]. This requires a one-to-one word alignment, which is generated by introduction of ε tokens and using an IBM1 translation table. We apply the *bidirectional* version of the translation model, which uses both forward and backward recurrency in order to take the *full source context* into account for each translation decision. The language models are set up with 300 nodes in both the projection and the hidden LSTM layer. For the BTM, we use 200 nodes in all layers, namely the forward and backward projection layers, the first hidden layers for both forward and backward processing and the second hidden layer, which joins the output of the directional hidden layers. The architecture of the BTM network is shown in Figure 1.

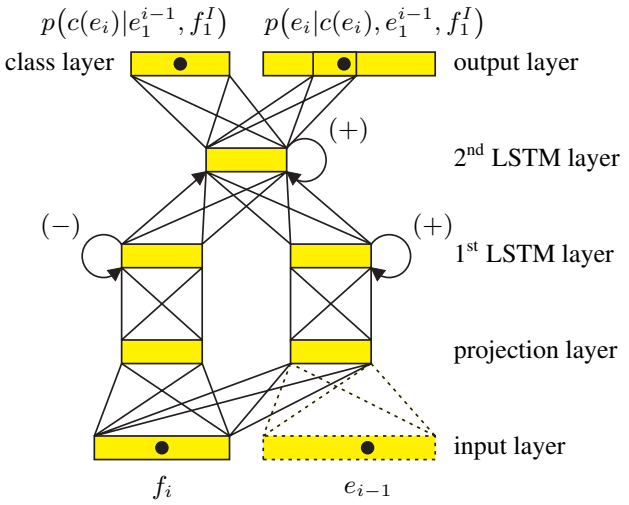

Figure 1: Architecture of the deep recurrent bidirectional translation model. By $(+)$ and $(-)$, we indicate a processing in forward and backward time directions, respectively. The inclusion of the dashed parts leads to a bidirectional *joint* model, which was not applied in this work. A single source projection matrix is used for the forward and backward branches.

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. As backoff language models, the baseline contains a general-domain LM, an in-domain LM and a word class LM (wcLM), which are described in Section 2.3. The hierarchical reordering model (HRM) is also contained in the baseline. Experimental results are given in Table 1. By maximum expected BLEU training of phrasal and lexical features, the baseline is improved by 0.7% BLEU absolute on `tst2010` and 1.5%

BLEU absolute on tst2011. Including the discriminatively trained reordering model yields further gains of 0.3 and 0.1 BLEU points. The recurrent language model gives us an additional 0.7 and 0.6 BLEU points and adding the recurrent translation model, we get 0.7% and 0.2% BLEU absolute on top. The observed improvements are confirmed on the blind evaluation set tst2014, on which the scores were computed by the workshop organizers. Thus, by applying only two general and language-independent methods, our state-of-the-art baseline is improved by 2.4% BLEU on tst2010, 3.5% BLEU on tst2011 and 2.7% BLEU on tst2014. Altogether compared to last year [21] our translation performance was increased by 1.7% BLEU and 1.5% TER absolute on tst2010.

3.2. German→English

Similar to English→French, 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. As backoff language models, the baseline contains a general-domain LM, an in-domain LM and a word class LM (wcLM), which are described in Section 2.3. The hierarchical reordering model (HRM) is also contained in the baseline. In a preprocessing step the German source was decompounded [22] and part-of-speech-based long-range verb reordering rules [23] were applied. In addition, we tuned our system on two different development sets (dev2010 and dev2012). Since the development set from 2010 is German translated from English talks, dev2012 contains manual transcriptions from German talks. As a real test set for the manual transcription is missing, we describe the results (Table 2) for the dev2010-tuned system in the following. By maximum expected BLEU training of phrasal and lexical features, the baseline is improved by 1.0% BLEU absolute on tst2010 and 1.6% BLEU absolute on tst2011. Including the discriminatively trained reordering model yields further gains of 0.4 and 0.2 BLEU points. The recurrent language model gives us an additional 0.7 and 1.1 BLEU points and adding the recurrent translation model, we get 0.7% and 0.6% BLEU absolute on top. Thus, by applying only two general and language-independent methods, our state-of-the-art baseline is improved by 2.8% BLEU on tst2010 and 3.5% BLEU on tst2011. Altogether compared to last year [21] our translation performance was increased by 1.8% BLEU and 2.2% TER absolute on tst2010. However, we submitted the system tuned on dev2012, which contains transcribed and translated German TED-X talks and is therefore more similar to the evaluation data. The improvements are similar to the system tuned on dev2010. Unfortunatelly, they do not carry over to the blind evaluation data tst2014 in the same magnitude, where we only observe a 0.8% gain over the baseline.

3.3. Spoken Language Translation (SLT)

RWTH participated in the English→French and German→English SLT tasks. For both language pairs, we reintroduced punctuation and case information before the actual translation similar to [24]. However, we 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 we used F_2-Score rather than BLEU or WER. More details can be found in [25].

Since punctuation predicting and recasing were applied before the actual translation, our translation systems could be kept completely unchanged and we were able to use our final systems from the MT track directly.

4. Conclusion

RWTH participated in two MT tracks and two SLT tracks of the IWSLT 2014 evaluation campaign. The baseline systems utilize our state-of-the-art phrase-based translation decoder and we were able to improve them by discriminative phrase training (+1.8 BLEU) and recurrent neural network models (+1.9 BLEU).

For the SLT track, the ASR output was enriched with punctuation and casing information by a hierarchical translation system tuned on F_2-Score.

All presented final systems are used in the EU-Bridge system combination [26].

5. 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.

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] 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.

[3] 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.

[4] R. Zens and H. Ney, "Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine

Table 1: Results for the English→French MT task. The scores on `tst2014` were computed by the task organizers.

system	dev2010		tst2010		tst2011		tst2014	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
SCSS 2013	30.0	53.8	33.7	48.0	-	-	-	-
SCSS baseline	29.8	54.5	33.0	48.5	39.0	41.5	33.8	46.1
+maxExpBleu phr+lex	30.5	54.2	33.7	48.2	40.5	40.4	-	-
+maxExpBleu RO	30.7	54.0	34.0	47.8	40.6	40.4	35.3	44.9
+RNN-LM	31.1	53.3	34.7	47.3	41.2	39.9	-	-
+RNN-BTM	31.8	52.6	35.4	46.5	42.5	39.0	36.5	43.8

Table 2: Results for the German→English MT task. The scores on `tst2014` were computed by the task organizers.

system	dev2010		dev2012		tst2010		tst2011		tst2012		tst2014	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
SCSS 2013	34.2*	45.8*			32.3	48.1	-	-	-	-	-	-
SCSS baseline	33.8*	46.5*	24.1	62.3	31.3	49.4	36.3	44.1	31.0	49.5	-	-
+maxExpBleu phr+lex	35.2*	45.2*	24.2	61.4	32.3	48.1	37.9	42.7	32.2	48.0	-	-
+maxExpBleu RO	35.4*	45.0*	24.5	61.2	32.7	47.9	38.1	42.5	32.7	47.6	-	-
+RNN-LM	35.8*	44.0*	25.6	59.8	33.4	46.9	39.2	41.4	32.9	47.1	-	-
+ RNN-BTM	36.3*	43.2*	26.2	58.7	34.1	45.9	39.8	40.6	33.5	46.0	25.0	56.1
SCSS baseline	33.0	46.0	26.8*	58.4*	30.7	48.8	37.0	42.9	30.8	48.5	24.8	55.6
+maxExpBleu phr+lex	34.0	44.4	27.1*	57.6*	32.5	46.9	38.3	41.4	32.4	46.6	-	-
+maxExpBleu RO	33.7	44.8	27.4*	57.7*	32.3	47.1	38.3	41.5	32.7	46.7	25.2	54.8
+RNN-LM	34.2	44.4	27.7*	56.6*	32.8	46.6	38.9	41.1	32.8	46.7	-	-
+RNN-BTM	34.7	44.2	27.8*	57.2*	33.2	46.5	39.4	40.7	33.2	46.3	25.6	54.6

:

Translation," in *International Workshop on Spoken Language Translation*, Honolulu, Hawaii, Oct. 2008, pp. 195–205.

[5] 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.

[6] B. Chen, R. Kuhn, G. Foster, and H. Johnson, "Unpacking and transforming feature functions: New ways to smooth phrase tables," in *MT Summit XIII*, Xiamen, China, Sept. 2011, pp. 269–275.

[7] 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.

[8] 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

[9] 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.

[10] 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.

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

[12] 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.

[13] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified Kneser-Ney language model estimation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013, pp. 690–696. [Online]. Available: http://kheafield.com/professional/edinburgh/estimate_paper.pdf

[14] R. Moore and W. Lewis, "Intelligent Selection of Language

Model Training Data," in *ACL (Short Papers)*, Uppsala, Sweden, July 2010, pp. 220–224.

[15] 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.

[16] 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.

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

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

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

[20] M. Sundermeyer, T. Alkhouli, J. Wuebker, and H. Ney, "Translation modeling with bidirectional recurrent neural networks," in *Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, Oct. 2014, pp. 14–25.

[21] J. Wuebker, S. Peitz, T. Alkhouli, J.-T. Peter, M. Feng, M. Freitag, and H. Ney, "The rwth aachen machine translation systems for iwslt 2013," in *International Workshop on Spoken Language Translation*, Heidelberg, Germany, Dec. 2013, pp. 88–93. [Online]. Available: http://workshop2013.iwslt.org/downloads/The_RWTH_Aachen_Machine_Translation_Systems_for_IWSLT_2013.pdf

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

[23] 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.

[24] 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.

[25] S. Peitz, M. Freitag, and H. Ney, "Better punctuation prediction with hierarchical phrase-based translation," in *International Workshop on Spoken Language Translation*, Lake Tahoe, CA, USA, Dec. 2014, to appear.

[26] M. Freitag, J. Wuebker, S. Peitz, H. Ney, M. Huck, A. Birch, N. Durrani, P. Koehn, M. Mediani, I. Slawik, J. Niehues, E. Cho, A. Waibel, N. Bertoldi, M. Cettolo, and M. Federico, "Combined spoken language translation," in *International Workshop on Spoken Language Translation*, Lake Tahoe, CA, USA, Dec. 2014, to appear.