The RWTH Aachen University English-Romanian Machine Translation System for WMT 2016

Jan-Thorsten Peter, Tamer Alkhouli, Andreas Guta and Hermann Ney
Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
<surname>@cs.rwth-aachen.de

Abstract

This paper describes the statistical machine translation system developed at RWTH Aachen University for the English→Romanian translation task of the *ACL 2016 First Conference on Machine Translation* (WMT 2016).

We combined three different state-of-the-art systems in a system combination: A phrase-based system, a hierarchical phrase-based system and an attention-based neural machine translation system. The phrase-based and the hierarchical phrase-based systems make use of a language model trained on all available data, a language model trained on the bilingual data and a word class language model. In addition, we utilized a recurrent neural network language model and a bidirectional recurrent neural network translation model for reranking the output of both systems. The attention-based neural machine translation system was trained using all bilingual data together with the back-translated data from the News Crawl 2015 corpora.

1 Introduction

We describe the statistical machine translation (SMT) systems developed by RWTH Aachen University for English→Romanian language pair for the evaluation campaign of WMT 2016. Combining several single machine translation engines has proven to be highly effective in previous submissions, e.g. (Freitag et al., 2013; Freitag et al., 2014a; Peter et al., 2015). We therefore used a similar approach for this evaluation. We trained individual systems using state-of-the-art phrase-based, hierarchical phrase-based translation engines, and attention-based recurrent neural networks ensemble. Each single system was optimized and the best systems were used in a system combination.

This paper is organized as follows. In Sections 2.2 through 2.5 we describe our translation software and baseline setups. Sections 2.6 describes the neural network models used in our translation systems. The attention based recurrent neural network ensemble is described in Section 2.7. Sections 2.8 explains the system combination pipeline applied on the individual systems for obtaining the combined system. Our experiments for each track are summarized in Section 3 and we conclude with Section 4.

2 SMT Systems

For the WMT 2016 evaluation campaign, the RWTH utilizes three different state-of-the-art translation systems:

- phrase-based
- hierarchical phrase-based
- attention based neural network ensemble

The phrase-based system is based on word alignments obtained with GIZA++ (Och and Ney, 2003). We use mteval from the Moses toolkit (Koehn et al., 2007) an TERCom to evaluate our systems on the BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) measures. All reported scores are case-sensitive and normalized.

2.1 Preprocessing

The preprocessing of the data was provided by LIMISI. The Romanian side was tokenized using their tokro toolkit (Allauzen et al., 2016 to appear). The English side was tokenized using the Moses toolkit (Koehn et al., 2007). Both sides were true cased with Moses.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 356–361,
Berlin, Germany, August 11-12, 2016.

2.2 Phrase-based Systems

Our phrase-based decoder (PBT) is the implementation of the *source cardinality synchronous search* (SCSS) procedure described in (Zens and Ney, 2008). It is freely available for non-commercial use in RWTH's open-source SMT toolkit, Jane 2.3[1] (Wuebker et al., 2012). Our baseline contains the following models: Phrase translation probabilities and lexical smoothing in both directions, word and phrase penalty, distance-based reordering model, n-gram target language models and enhanced low frequency feature (Chen et al., 2011), a hierarchical reordering model (HRM) (Galley and Manning, 2008), and a high-order word class language model (wcLM) (Wuebker et al., 2013) trained on all monolingual data. The phrase table is trained on all bilingual data. Additionally we add synthetic parallel data as described in Section 2.4. Two different neural network models (cf. Sections 2.6) are applied in reranking. The parameter weights are optimized with MERT (Och, 2003) towards the BLEU metric.

2.3 Hierarchical Phrase-based System

The open source translation toolkit Jane 2.3 (Vilar et al., 2010) is also used for our hierarchical setup. Hierarchical phrase-based translation (HPBT) (Chiang, 2007) induces a weighted synchronous context-free grammar from parallel text. Additional to the contiguous *lexical* phrases, as used in PBT, *hierarchical* phrases with up to two gaps are extracted. Our baseline model contains models with phrase translation probabilities and lexical smoothing probabilities in both translation directions, word and phrase penalty. It also contains binary features to distinguish between hierarchical on non-hierarchical phrases, the glue rule, and rules with non-terminals at the boundaries. The enhanced low frequency feature (Chen et al., 2011) and the same n-gram language models as described in our PBT system are also used. We utilize the cube pruning algorithm (Huang and Chiang, 2007) for decoding. Neural networks are applied in reranking similar to the PBT system and the parameter weights are also optimized with MERT (Och, 2003) towards the BLEU metric.

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

2.4 Synthetic Source Sentences

The training data contains around 600k bilingual sentence pairs. To increase the amount of usable training data for the phrase-based and the neural machine translation systems we translated part of the monolingual training data back to English in a similar way as described by (Bertoldi and Federico, 2009) and (Sennrich et al., 2016 to appear).

We created a simple baseline phrase-based system for this task. All bilingual data is used to extract the phrase table and the system contains one language model which uses the English side of the bilingual data combined with the English News Crawl 2007-2015, News Commentary and News Discussion data.

This provides us with nearly 2.3M additional parallel sentences for training. The phrase-based system as well as the attention-based neural network system are trained with this additional data.

2.5 Backoff Language Models

Both phrase-based and hierarchical translation systems use three backoff language models (LM) that are estimated with the KenLM toolkit (Heafield et al., 2013) and are integrated into the decoder as separate models in the log-linear combination: A full 4-gram LM (trained on all data), a limited 5-gram LM (trained only on indomain data), and a 7-gram word class language model (wcLM). All of them use interpolated Kneser-Ney smoothing. 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 described in (Wuebker et al., 2013) to compute the wcLM on the same data as the large LM.

2.6 Recurrent Neural Network Models

Our systems apply reranking on 1000-best lists using recurrent language and translation models. We use the long short-term memory (LSTM) architecture for recurrent layers (Hochreiter and Schmidhuber, 1997; Gers et al., 2000; Gers et al., 2003). The models have a class-factored output layer (Goodman, 2001; Morin and Bengio, 2005) to speed up training and evaluation. The class layer consists of 2000 word classes. The LSTM recurrent neural network language model (RNN-LM) (Sundermeyer et al., 2012) uses a vocabulary of 143K words. It is trained on the concatenation of the English side of the parallel data and the News

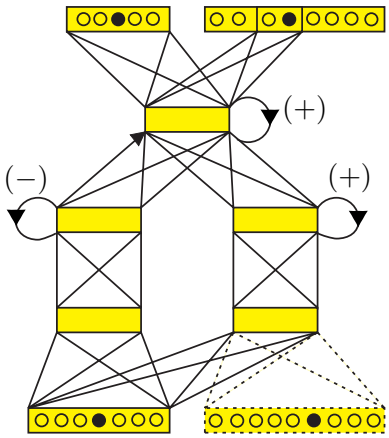

Figure 1: The architecture of the deep bidirectional joint model. By $(+)$ and $(-)$, we indicate a processing in forward and backward time directions, respectively. The dashed part indicates the target input. The model has a class-factored output layer.

Crawl 2015 corpus, amounting to 2.9M sentences (70.7M running words). We use one projection layer, and 3 stacked LSTM layers, with 350 nodes each.

In addition to the RNN-LM, we apply the deep bidirectional joint model (BJM) described in (Sundermeyer et al., 2014a) in 1000-best reranking. As the model depends on the complete alignment path, this variant cannot be applied directly in decoding (Alkhouli et al., 2015). The model assumes a one-to-one alignment between the source and target sentences. This is generated by assigning unaligned source and target words to $\epsilon_{unaligned}$ tokens that are added to the source and target vocabularies. In addition the source and target vocabularies are extended to include $\epsilon_{aligned}$ tokens, which are used to break down multiply-aligned source and target words using the IBM-1 translation tables. For more details we refer the reader to (Sundermeyer et al., 2014a).

The BJM has a projection layer, and computes a forward recurrent state encoding the source and target history, a backward recurrent state encoding the source future, and a third LSTM layer to combine them. The architecture is shown in Figure 1. All layers have 350 nodes. The model was trained on 604K sentence pairs, having 15.4M and 15.7M source and target words respectively. The has respectively 33K and 55K source and target vocabulary.

The neural networks were implemented using

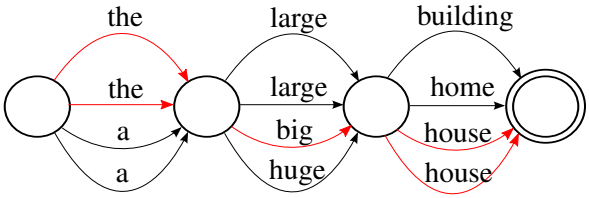

Figure 2: System A: *the large building*; System B: *the large home*; System C: *a big house*; System D: *a huge house*; Reference: *the big house*.

an extension of the RWTHLM toolkit (Sundermeyer et al., 2014b).

2.7 Attention Based Recurrent Neural Network

The second system provided by the RWTH is an attention-based recurrent neural network (NMT) similar to (Bahdanau et al., 2015). We use an implementation based on Blocks (van Merriënboer et al., 2015) and Theano (Bergstra et al., 2010; Bastien et al., 2012).

The network uses the 30K most frequent words on the source and target side as input vocabulary. The decoder and encoder word embeddings are of size 620, the encoder uses a bidirectional layer with 1024 GRUs (Cho et al., 2014) to encode the source side. A layer with 1024 GRUs is used by the decoder.

The network is trained for up to 300K iterations with a batch size of 80. The network was evaluated every 10000 iterations and the best network on the newsdev2016/1 dev set was selected.

The synthetic training data is used as described in Section 2.4 to create additional parallel training data. The new data is weighted by using the News Crawl 2015 corpus (2.3M sentences) once, the Europarl corpus (0.4M sentences) twice and the SETimes2 corpus (0.2M sentences) three times. We use an ensemble of 4 networks, all with the same configuration and training settings. If the neural network creates unknown word the source word where the strongest attention weight points to is copied to the target side. We did not use any regularization as dropout or Gaussian noise.

2.8 System Combination

System combination is applied to produce consensus translations from multiple hypotheses which are obtained from different translation approaches. The consensus translations outperform the individual hypotheses in terms of translation quality.

Table 1: Results of the individual systems for the English→Romanian task. BLEU and TER scores are case-sensitive and given in %.

Individual Systems	newsdev2016/1		newsdev2016/2		newstest2016	
	BLEU	TER	BLEU	TER	BLEU	TER
Phrase-Based	23.7	60.3	27.8	54.7	24.4	58.9
+ additional parallel data	24.3	59.4	29.2	53.0	25.0	58.2
+ NNs	26.0	55.9	31.4	50.7	26.0	56.0
Hierarchical	23.8	60.6	27.9	54.7	24.5	59.0
+ NNs	26.1	56.4	29.7	52.4	25.5	57.1
Attention Network	20.9	63.1	22.7	58.7	21.2	61.5
+ additional parallel data	23.4	59.4	27.6	52.7	24.0	58.0
+ ensemble	25.6	55.0	30.7	48.8	26.1	54.9
System Combination	27.6	55.0	31.7	50.3	26.9	55.4

A system combination implementation which has been developed at RWTH Aachen University (Freitag et al., 2014b) is used to combine the outputs of different engines.

The first step in system combination is generation of confusion networks (CN) from I input translation hypotheses. We need pairwise alignments between the input hypotheses, and the alignments are obtained by METEOR (Banerjee and Lavie, 2005). The hypotheses are then reordered to match a selected skeleton hypothesis in terms of word ordering. We generate I different CNs, each having one of the input systems as the skeleton hypothesis, and the final lattice will be the union of all I generated CNs. In Figure 2 an example of a confusion network with $I = 4$ input translations is depicted. The decoding of a confusion network is finding the shortest path in the network. Each arc is assigned a score of a linear model combination of M different models, which include word penalty, 3-gram language model trained on the input hypotheses, a binary primary system feature that marks the primary hypothesis, and a binary voting feature for each system. The binary voting feature for a system is 1 iff the decoded word is from that system, and 0 otherwise. The different model weights for system combination are trained with MERT.

3 Experimental Evaluation

All three systems use the same preprocessing as described in Section 2.1. The phrase-based system in its baseline configuration was improved by 0.6 BLEU and 0.7 TER points on newstest2016 by adding the synthetic data as described in Section 2.4. The neural networks (Section 2.6 improve the network by another 1.0 BLEU and 2.2 TER.

Table 2: Comparing the systems against each other by computing the BLEU and TER score on the newstest2016. Each system is used as reference once, the reported value is the average between both which makes these value symmetrical. The upper half lists BLEU scores, the lower half TER scores. All values are given in %.

	PBT	HPBT	NMT	Average
PBT	-	62.6	51.1	56.9
HPBT	24.9	-	47.5	55.1
NMT	31.8	34.8	-	49.3
Average	28.3	29.8	33.3	

The neural networks also improve the hierarchical phrase-based system by 1.0 BLEU and 2.9 TER. We did not try to add the synthetic data to the hierarchical system.

Adding the synthetic data to the NMT system improve the baseline system by 3.8 BLEU and 3.5 TER. An ensemble of four similarly trained networks gives an additional improvement of 2.1 BLEU and in 3.1 TER.

The final step was to combine all three systems using the system combination (Section 2.8) which added another 0.8 BLEU points on top of the neural network system, but caused a small degradation in TER by 0.5 points.

The lower BLEU and higher TER score in Table 2 for the NMT system show that the translations created by it differ more from the PBT and HPBT system then there translation between each other.

4 Conclusion

RWTH participated with a system combination on the English→Romanian WMT 2016 evaluation campaign. The system combination included three different state-of-the-art systems: A phrase-based, a hierarchical phrase-based and a stand alone attention-based neural network system. The phrase-based and the hierarchical phrase-based systems where both supported by a neural network LM and BJM. Synthetic data was used to improve the amount of parallel data for the PBT and the NMT system.

We achieve a performance of 26.9 BLEU and 55.4 TER on the newstest2016 test set.

Acknowledgments

This paper has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 645452 (QT21).

References

Tamer Alkhouli, Felix Rietig, and Hermann Ney. 2015. Investigations on phrase-based decoding with recurrent neural network language and translation models. In *EMNLP 2015 Tenth Workshop on Statistical Machine Translation*, pages 294–303, Lisbon, Portugal, September.

Alexandre Allauzen, Lauriane Aufrant, Franck Burlot, Elena Knyazeva, Thomas Lavergne, and François Yvon. 2016, to appear. LIMSI@WMT'16 : Machine translation of news. In *Proceedings of the Eleventh Workshop on Statistical Machine Translation*, Berlin, Germany, August. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, May.

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

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. 2012. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June. Oral Presentation.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 182–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Boxing Chen, Roland Kuhn, George Foster, and Howard Johnson. 2011. Unpacking and Transforming Feature Functions: New Ways to Smooth Phrase Tables. In *MT Summit XIII*, pages 269–275, Xiamen, China, September.

D. Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

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. Cetolo, and M. Federico. 2013. EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 128–135, Heidelberg, Germany, December.

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. 2014a. EU-BRIDGE MT: Combined Machine Translation. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 105–113, Baltimore, MD, USA, June.

Markus Freitag, Matthias Huck, and Hermann Ney. 2014b. Jane: Open Source Machine Translation System Combination. In *Proc. of the Conf. of the European Chapter of the Assoc. for Computational Linguistics (EACL)*, pages 29–32, Gothenberg, Sweden, April.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856, Stroudsburg, PA, USA. Association for Computational Linguistics.

Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with LSTM. *Neural computation*, 12(10):2451–2471.

Felix A. Gers, Nicol N. Schraudolph, and Jürgen Schmidhuber. 2003. Learning precise timing with lstm recurrent networks. *The Journal of Machine Learning Research*, 3:115–143.

Joshua Goodman. 2001. Classes for fast maximum entropy training. *CoRR*, cs.CL/0108006.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantine, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. pages 177–180, Prague, Czech Republic, June.

Frederic Morin and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 246–252. Society for Artificial Intelligence and Statistics.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Jan-Thorsten Peter, Farzad Toutounchi, Stephan Peitz, Parnia Bahar, Andreas Guta, and Hermann Ney. 2015. The rwth aachen german to english mt system for iwslt 2015. In *International Workshop on Spoken Language Translation*, pages 15–22, Da Nang, Vietnam, December.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016, to appear. Improving neural machine translation models with monolingual data. August.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. 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*, pages 223–231, Cambridge, Massachusetts, USA, August.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM Neural Networks for Language Modeling. In *Interspeech*, Portland, OR, USA, September.

Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney. 2014a. Translation Modeling with Bidirectional Recurrent Neural Networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 14–25, Doha, Qatar, October.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2014b. rwthlm - the rwth aachen university neural network language modeling toolkit. In *Interspeech*, pages 2093–2097, Singapore, September.

Bart van Merriënboer, Dzmitry Bahdanau, Vincent Dumoulin, Dmitriy Serdyuk, David Warde-Farley, Jan Chorowski, and Yoshua Bengio. 2015. Blocks and fuel: Frameworks for deep learning. *CoRR*, abs/1506.00619.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2010. Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 262–270, Uppsala, Sweden, July.

Joern Wuebker, Matthias Huck, Stephan Peitz, Malte Nuhn, Markus Freitag, Jan-Thorsten Peter, Saab Mansour, and Hermann Ney. 2012. Jane 2: Open source phrase-based and hierarchical statistical machine translation. In *International Conference on Computational Linguistics*, pages 483–491, Mumbai, India, December.

Joern Wuebker, Stephan Peitz, Felix Rietig, and Hermann Ney. 2013. Improving statistical machine translation with word class models. In *Conference on Empirical Methods in Natural Language Processing*, pages 1377–1381, Seattle, WA, USA, October.

Richard Zens and Hermann Ney. 2008. Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation. In *International Workshop on Spoken Language Translation*, pages 195–205, Honolulu, Hawaii, October.