The RWTH Aachen German-English Machine Translation System for WMT 2015

Jan-Thorsten Peter, Farzad Toutounchi, Joern Wuebker 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 German→English translation task of the *EMNLP 2015 Tenth Workshop on Statistical Machine Translation* (WMT 2015). A phrase-based machine translation system was applied and augmented with hierarchical phrase reordering and word class language models. Further, we ran discriminative maximum expected BLEU training for our system. In addition, we utilized multiple feed-forward neural network language and translation models and a recurrent neural network language model for reranking.

1 Introduction

For the WMT 2015 shared translation task[1], RWTH utilized a state-of-the-art phrase-based translation system. We participated in the German→English translation task. The system included a hierarchical reordering model, a word class (cluster) language model, and discriminative maximum expected BLEU training. Further, we reranked the nbest lists produced by our system with three feed-forward neural network models and a recurrent neural language model.

This paper is structured as follows: First, we briefly describe our preprocessing pipeline for the language pair German→English in Section 2, which is based on our 2014 pipeline. Next, morpho-syntactic analysis for preprocessing the data is described in Section 2.3. Different alignment methods are discussed in Section 3. In Section 4, we present a summary of all methods used in our submission. More details are given about the language models (Section 4.2), maximum expected BLEU training (Section 4.4), the hierarchical reordering model (Section 4.5), feed-forward neural network training (Section 4.6), and recurrent neural network language model (Section 4.7). Experimental results are discussed in Section 5. We conclude the paper in Section 6.

[1] http://www.statmt.org/wmt15/translation-task.html

2 Preprocessing

In this section we briefly describe our preprocessing pipeline, which is a modification of our WMT 2014 German→English preprocessing pipeline (Peitz et al., 2014).

2.1 Categorization

We worked on the categorization of the digits and written numbers for the translation task. All written numbers were categorized. As the training data and also the test sets contain several errors for numbers in the source as well as in the target part, we put effort into producing correct English numbers. In addition, ',' and '.' marks were inverted in most cases, as in German the former mark is the decimal mark and the latter is the thousand separator.

2.2 Remove Foreign Languages

The WMT German→English Common Crawl corpora contains bilingual sentence pairs with non-German source or non-English target sentences. By using an ASCII filtering, we removed all sentences with more than 5% non-ASCII characters from the Common Crawl corpus. Chinese, Arabic and Russian are among the languages which can be easily filtered by deleting the sentences containing too many non-ASCII words. Our experiments showed that the translation quality does not change by removing sentences with wrong languages. Nevertheless, this method reduced the training data size and also the vocabulary size without introducing any degradation in translation

Table 1: Comparison of a simple GIZA++ alignment vs. merging multiple alignments. Even though the multiple alignment approach did not improve the GIZA++ alignment for the baseline system, it improved translation quality in combination with a neural network joint model (NNJM). BLEU and TER are given in percentage.

	newstest2011		**newstest2012**		**newstest2013**		**newstest2014**	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
GIZA++	23.1	58.8	23.7	58.2	26.5	54.7	25.9	54.2
+ NNJM	23.3	58.4	24.0	57.7	26.6	54.3	26.2	53.7
Multiple alignment	23.0	58.9	23.8	58.2	26.6	54.6	25.9	54.2
+ NNJM	23.3	58.4	24.1	57.8	27.0	54.3	26.3	53.8

quality. Further, this method prevents us from generating words from these languages.

2.3 Compound Splitting and POS-based Word Reordering

We reduced the source vocabulary size for the German→English translation and split the German compound words with the frequency-based method described in Koehn and Knight (2003). To reduce translation complexity, we employed the long-range part-of-speech based reordering rules proposed by Popović and Ney (2006). In this regard, we did no further morphological analysis in our preprocessing pipeline.

3 Alignment

We experimented with creating multiple alignments and merging them via a majority vote. For the majority voting to work in a meaningful way we need obviously more than two different alignments. A larger number of alignments gives us more confidence that the alignment points are correct.

To create these different alignments, we used `fast_align` (Dyer et al., 2013) and two implementations of GIZA++ (Och and Ney, 2003). The alignment was trained in both source to target direction and target to source direction. To double the number of alignments, we trained each setup also with a reverse ordered source side and reversed it back after the alignment process finished (Freitag et al., 2013). Using a reversed source side usually creates a different alignment since the word order influences the results of `fast_align` and GIZA++. This gave us a total of 12 different alignments (three toolkits $\times$ two translation directions $\times$ two source side direction). These alignments were merged by keeping all alignment points generated by at least 5 of the methods.

We compared this setup with an alignment generated by GIZA++. The voting setup did not improve directly on the baseline system as shown in Table 1. However, in combination with a feed-forward neural network joint model (Section 4.6) the results on `newstest2013` improved by 0.4% BLEU after reranking. We stuck in the following experiments to the multiple alignments approach.

4 Translation System

In this evaluation, we used the open source machine translation toolkit *Jane*[2] (Vilar et al., 2012; Wuebker et al., 2012). This open-source toolkit was developed at the RWTH Aachen University and includes a phrase-based decoder used in all of our experiments.

4.1 Phrase-based System

Our phrase based decoder includes an implementation of the source cardinality synchronous search procedure described in Zens and Ney (2008). We used the standard set of models with phrase translation probabilities, lexical smoothing in both directions, word and phrase penalty, distance-based distortion model, a 4-gram target language model and enhanced low frequency feature (Chen et al., 2011). Additional models used in this evaluation were the hierarchical reordering model (*HRM*) (Galley and Manning, 2008) and a word class language model (*wcLM*) (Wuebker et al., 2013). The parameter weights were optimized with minimum error rate training (MERT) (Och, 2003). The op-

[2] `http://www.hltpr.rwth-aachen.de/jane/`

timization criterion was BLEU (Papineni et al., 2002).

4.2 Language Models

We used a 4-gram language model trained on the target side of the bilingual data, $\frac{1}{2}$ of the Shuffled News Crawl corpus, $\frac{1}{2}$ of the 10^9 French-English corpus and $\frac{1}{4}$ of the LDC Gigaword Fifth Edition corpus. The monolingual data selection was based on cross-entropy difference as described in Moore and Lewis (2010). For this language model, we trained separate language models using SRILM for each corpus, which were then interpolated. The interpolation weights are tuned by minimizing the perplexity of the interpolated model on the development data. In addition, a word class language model was utilized. We trained 200 classes on the target side of the bilingual training data (Brown et al., 1992; Och, 1999). We used the same data as the 4-gram language model for training a 7-gram wcLM. Furthermore, we also trained a single unpruned language model on the concatenation of all monolingual data using KenLM, which was used as an extra model in our final experiments. All language models used interpolated Kneser-Ney smoothing.

4.3 Evaluation

All setups were evaluated with *MultEval* (Clark et al., 2011). To evaluate our models, we used the average of three MERT optimization runs for case sensitive BLEU (Papineni et al., 2002) and case insensitive TER[3] (Snover et al., 2006).

4.4 Maximum Expected BLEU Training

In our baseline translation system the phrase tables were extracted from word alignments and the probabilities were estimated as relative frequencies, which is still the state-of-the-art for many standard SMT systems. For the WMT 2015 German→English task, we applied discriminative maximum expected BLEU training as described by Wuebker et al. (2015). The expected BLEU objective function is optimized with the resilient backpropagation algorithm (RPROP) (Riedmiller and Braun, 1993). Similar to He and Deng (2012), the objective function is computed on n-best lists (here: $n = 100$) generated by the translation decoder. To avoid over-fitting due to spurious

[3]TER is always evaluated in case insensitive form by MultEval.

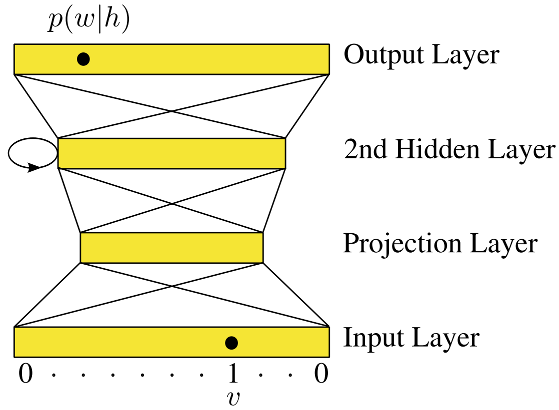

Figure 1: LM neural network

segmentations, we apply a leave-one-out heuristic (Wuebker et al., 2010) during the n-best list generation step. Using these n-best lists, we iteratively trained the phrasal and lexical feature sets, denoted as (a) and (b) in Wuebker et al. (2015). Each of the two feature types are condensed into a single model within the log-linear model combination. After every five iterations we ran MERT, and finally selected the iteration performing best on `newstest2013`. In this work, we used a subset of the training data to generate the n-best lists, namely the concatenation of `newstest2008` through `newstest2010` and the News-Commentary corpus.

4.5 Hierarchical Reordering Model

In Galley and Manning (2008), a hierarchical reordering model for phrase-based machine translation was introduced. The model scores *monotone*, *swap*, and *discontinuous* phrase orientations in the manner of the one presented by Tillmann (2004). The orientation classes are determined based on phrase *blocks*, which can subsume multiple phrase pairs and are computed with an SR-parser. The model has proven effective in previous evaluations. As the word order is more flexible in German compared to English, we expected that an additional reordering model could improve the translation quality.

4.6 Feed-Forward Neural Network Training

We used three feed-forward neural network (*FFNN*) models with a similar structure as the network models used by Devlin et al. (2014) and Le et al. (2012). All networks were trained with different input features:

- Translation Model (TM), the 5 source words around the alignment source word

Table 2: Results for the German→English translation task. The results are the average of three optimization runs. `newstest2011` and `newstest2012` were used as development data. The submission system used all models and the best optimization run on the development data. BLEU and TER are given in percentage.

	newstest2011		newstest2012		newstest2013		newstest2014	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
Baseline	23.0	58.9	23.8	58.2	26.6	54.6	25.9	54.2
+ max. exp. BLEU	23.1	58.6	24.0	57.8	26.8	54.4	26.2	53.9
+ updated LM	23.2	58.7	24.0	57.9	26.8	54.3	26.3	53.7
+ unpruned LM	23.2	59.0	24.1	58.1	26.9	54.6	26.6	54.0
+ 3 × FFNN	23.7	58.4	24.5	57.7	27.4	54.0	27.1	53.3
+ LSTM	23.8	58.4	24.7	57.4	27.5	53.8	27.1	53.2
Submission System	24.1	57.6	25.0	56.5	28.1	52.9	27.6	52.3

- Language Model (LM), the 7 last words on the target side
- Joint Model (JM), the 5 source words around the alignment source word and the 4 last words on the target side

The TM and LM were trained with two hidden layers (1000 and 500 nodes) while the JM contained three hidden layers with 2000 nodes each. The output layer was in all cases a softmax layer with a short list of 10000. All remaining words were clustered into 1000 classes and their class probabilities were predicted. The neural networks were applied to rerank 1000-best lists.

4.7 Recurrent Neural Network Language Model

In addition to the feed-forward neural network model we employed a recurrent neural network model. The recurrency was handled with the long short-term memory (*LSTM*) architecture (Hochreiter and Schmidhuber, 1997) and we used a class-factored output layer for increased efficiency as described in Sundermeyer et al. (2012). The topology of the network is illustrated in Figure 1. All neural network models were trained on the bilingual data with 2000 word classes. The language models were set up with 500 nodes in both the projection layer and the hidden LSTM layer. The recurrent network models were applied together with the feed-forward models to rerank 1000-best lists.

5 Setup

We trained the phrase-based system on all available bilingual training data. The preprocessed bilingual corpus contained around 4 million sentences. The preprocessed data contained a source vocabulary size of 814K and a target vocabulary size of 733K.

We used the target side of the bilingual data along with the monolingual corpora for training the language models. First, we started using our old language models from our WMT 2014 setup as baseline. Then we updated our system to the new language models trained according to Section 4.2. All results are reported as average of three optimization runs.

5.1 Experimental Results

The results of the phrase-based system are summarized in Table 2. It was tuned on the concatenation of `newstest2011` and `newstest2012`.

The phrase-based baseline system, which included the hierarchical reordering model (Galley and Manning, 2008) and a word class language model (*wcLM*) (Wuebker et al., 2013), reached a performance of 25.9% BLEU on `newstest2014`. Maximum expected BLEU training selected on `newstest2013` improved the results on `newstest2014` by 0.3% BLEU absolute.

There was improvement of 0.1% in BLEU on `newstest2014` by replacing the old language models from WMT 2014 with an updated general 4-gram LM and word class LM. Furthermore, adding an extra unpruned language model trained on the concatenation of the monolingual data improved the results on newstest2014 by 0.3% BLEU.

Adding three feed-forward neural network models yielded an improvement of 0.5% BLEU on `newstest2013` and `newstest2014`. Adding the LSTM language model improved the TER by an additional 0.1% on `newstest2014` and by 0.2% on `newstest2013`.

The submission system used all models and we chose the best optimization run on the development data. This optimization run by itself was 0.5% BLEU stronger on `newstest2014` compared to the average across three optimization runs which included this run.

6 Conclusion

For the participation in the WMT 2015 shared translation task, RWTH experimented with a phrase-based translation system. For this approach, we applied a hierarchical phrase reordering model and a word class language model. `fast_align` and two versions of GIZA++ were used for training word alignments, and a voting setup was implemented, which improved the results in combination with neural network models. We also employed discriminative maximum expected BLEU training. Additionally, we utilized feed-forward and recurrent neural networks models for our phrase-based system, which improved the performance. Furthermore, we adapted our preprocessing pipeline based on our WMT 2014 setup. Filtering the corpus for non-ASCII letters gave us lower vocabulary sizes for both source and target side without loss in performance.

Acknowledgments

This paper has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 645452 (QT21). We further thank the anonymous reviewers for their valuable comments.

References

Peter F. Brown, Vincent J. Della Pietra, P. V. deSouza deSouza, J. C. Lai, and Robert L. Mercer. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479.

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.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 176–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland, June. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparametrization of ibm model 2. In *Proceedings of the NAACL 7th Workshop on Syntax, Semantics and Structure in Statistical Translation*, Atlanta, Georgia, USA, June.

Markus Freitag, Minwei Feng, Matthias Huck, Stephan Peitz, and Hermann Ney. 2013. Reverse word order models. In *Machine Translation Summit*, pages 159–166, Nice, France, September.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 847–855, Honolulu, Hawaii, USA, October.

Xiaodong He and Li Deng. 2012. Maximum Expected BLEU Training of Phrase and Lexicon Translation Models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 292–301, Jeju, Republic of Korea, Jul.

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

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of European Chapter of the ACL (EACL 2009)*, pages 187–194.

Hai-Son Le, Alexandre Allauzen, and François Yvon. 2012. Continuous space translation models with neural networks. pages 39–48, Montréal, Canada, June. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent Selection of Language Model Training Data. In *ACL (Short Papers)*, pages 220–224, Uppsala, Sweden, July.

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. 1999. An Efficient Method for Determining Bilingual Word Classes. In *Proc. 9th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics*, pages 71–76, Bergen, Norway, June.

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.

Stephan Peitz, Joern Wuebker, Markus Freitag, and Hermann Ney. 2014. The RWTH Aachen German-English Machine Translation System for WMT 2014. In *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, June.

Maja Popović and Hermann Ney. 2006. POS-based Word Reorderings for Statistical Machine Translation. In *International Conference on Language Resources and Evaluation*, pages 1278–1283, Genoa, Italy, May.

Martin Riedmiller and Heinrich Braun. 1993. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591.

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.

Christoph Tillmann. 2004. A Unigram Orientation Model for Statistical Machine Translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 101–104, Boston, MA, USA.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2012. Jane: an advanced freely available hierarchical machine translation toolkit. *Machine Translation*, 26(3):197–216, September.

Joern Wuebker, Arne Mauser, and Hermann Ney. 2010. Training phrase translation models with leaving-one-out. In *Proceedings of the 48th Annual Meeting of the Assoc. for Computational Linguistics*, pages 475–484, 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, USA, October.

Joern Wuebker, Sebastian Muehr, Patrick Lehnen, Stephan Peitz, and Hermann Ney. 2015. A comparison of update strategies for large-scale maximum expected bleu training. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1516–1526, Denver, CO, USA, May.

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, USA, October.