The RWTH Aachen German-English Machine Translation System for WMT 2014

Stephan Peitz, Joern Wuebker, Markus Freitag 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 (SMT) systems developed at RWTH Aachen University for the German→English translation task of the *ACL 2014 Eighth Workshop on Statistical Machine Translation* (WMT 2014). Both hierarchical and phrase-based SMT systems are applied employing hierarchical phrase reordering and word class language models. For the phrase-based system, we run discriminative phrase training. In addition, we describe our preprocessing pipeline for German→English.

1 Introduction

For the WMT 2014 shared translation task[1] RWTH utilized state-of-the-art phrase-based and hierarchical translation systems. First, we describe our preprocessing pipeline for the language pair German→English in Section 2. Furthermore, we utilize morpho-syntactic analysis to preprocess the data (Section 2.3). In Section 3, we give a survey of the employed systems and the basic methods they implement. More details are given about the discriminative phrase training (Section 3.4) and the hierarchical reordering model for hierarchical machine translation (Section 3.5). Experimental results are discussed in Section 4.

2 Preprocessing

In this section we will describe the modification of our preprocessing pipeline compared to our 2013 WMT German→English setup.

2.1 Categorization

We put some effort in building better categories for digits and written numbers. All written numbers were categorized. In 2013 they were just handled as normal words which leads to a higher number of out-of-vocabulary words. For German→English, in most cases for numbers like '3,000' or '2.34' the decimal mark ',' and the thousands separator '.' has to be inverted. 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 more effort into producing correct English numbers.

2.2 Remove Foreign Languages

The WMT German→English corpus contains some bilingual sentence pairs with non-German source or/and non-English target sentences. For this WMT translation task, we filtered all non-matching language pairs (in terms of source language German and target language English) from our bilingual training set.

First, we filtered languages which contain non-ascii characters. For example Chinese, Arabic or Russian can be easily filtered when deleting sentences which contain more than 70 percent non-ascii words. The first examples of Table 1 was filtered due to the fact, that the source sentence contains too many non-ascii characters.

In a second step, we filtered European languages containing ascii characters. We used the WMT monolingual corpora in Czech, French, Spanish, English and German to filter these languages from our bilingual data. We could both delete a sentence pair if it contains a wrong source language or a wrong target language. That is the reason why we even search for English sentences in the source part and for German sentences in the target part. For each language, we built a word count of all words in the monolingual data for each language separately. We removed punctuation which are no indicator of a language. In our experiments, we only considered words with frequency higher than 20 (e.g. to ignore names). Given the word frequency, we removed a bilingual

[1] http://www.statmt.org/wmt14/translation-task.html

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 157–162,
Baltimore, Maryland USA, June 26–27, 2014. ©2014 Association for Computational Linguistics

Table 1: Examples of sentences removed in preprocessing.

	Example
remove non-ascii symbols	高效的技以抵消影响 . zum Bericht Añoveros Trías de Bes
remove wrong languages from target	Honni soit qui mal y pense ! as you yourself have said : travailler plus pour gagner plus
remove wrong languages from source	je déclare interrompue la session du Parlement européen . Quelle der Tabelle : “ what Does the European Union do ? ”

sentence pair from our training data if more than 70 percent of the words had a higher count in a different language then the one we expected. In Table 1 some example sentences, which were removed, are illustrated.

In Table 2 the amount of sentences and the corresponding vocabulary sizes of partial and totally cleaned data sets are given. Further we provide the number of out-of-vocabulary words (OOVs) for *newstest2012*. The vocabulary size could be reduced by ∼130k words for both source and target side of our bilingual training data while the OOV rate kept the same. Our experiments showed, that the translation quality is the same with or without removing wrong sentences. Nevertheless, we reduced the training data size and also the vocabulary size without any degradation in terms of translation quality.

2.3 Morpho-syntactic Analysis

In order to reduce the source vocabulary size for the German→English translation further, the German text is preprocessed by splitting German compound words with the frequency-based method described in (Koehn and Knight, 2003). To reduce translation complexity, we employ the long-range part-of-speech based reordering rules proposed by Popović and Ney (2006).

3 Translation Systems

In this evaluation, we employ phrase-based translation and hierarchical phrase-based translation. Both approaches are implemented in *Jane* (Vilar et al., 2012; Wuebker et al., 2012), a statistical machine translation toolkit which has been developed at RWTH Aachen University and is freely available for non-commercial use.[2] In the newest internal version, we use the KenLM Language Model Interface provided by (Heafield, 2011) for both decoders.

3.1 Phrase-based System

In the phrase-based decoder (source cardinality synchronous search, *SCSS*, Wuebker et al. (2012)), we use the standard set of models with phrase translation probabilities and lexical smoothing in both directions, word and phrase penalty, distance-based distortion model, an n-gram target language model and three binary count features. Additional models used in this evaluation are the hierarchical reordering model (*HRM*) (Galley and Manning, 2008) and a word class language model (*wcLM*) (Wuebker et al., 2013). The parameter weights are optimized with minimum error rate training (MERT) (Och, 2003). The optimization criterion is BLEU (Papineni et al., 2002).

3.2 Hierarchical Phrase-based System

In hierarchical phrase-based translation (Chiang, 2007), 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 hierarchical systems (Vilar et al., 2010; Huck et al., 2012) 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, three binary count features, and an n-gram language model. We utilize the cube pruning algorithm for decoding (Huck et al., 2013a) and optimize the model weights with MERT. The optimization criterion is BLEU.

3.3 Other Tools and Techniques

We employ GIZA++ (Och and Ney, 2003) to train word alignments. The two trained alignments are heuristically merged to obtain a symmetrized word alignment for phrase extraction. All lan-

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

Table 2: Corpus statistics after each filtering step and compound splitting.

		Vocabulary		OOVs
	Sentences	German	English	newstest2012
Preprocessing 2013	4.19M	1.43M	784K	1019
Preprocessing 2014	4.19M	1.42M	773K	1018
+ remove non-ascii symbols	4.17M	1.36M	713K	1021
+ remove wrong languages from target	4.15M	1.34M	675K	1027
+ remove wrong languages from source	4.08M	1.30M	655K	1039
+ compound splitting	4.08M	652K	655K	441

guage models (*LMs*) are created with the SRILM toolkit (Stolcke, 2002) or with the KenLM language model toolkit (Heafield et al., 2013) and are standard 4-gram LMs with interpolated modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998). We evaluate in truecase with BLEU and TER (Snover et al., 2006).

3.4 Discriminative Phrase Training

In our baseline translation systems the phrase tables are created by a heuristic extraction from word alignments and the probabilities are estimated as relative frequencies, which is still the state-of-the-art for many standard SMT systems. Here, we applied a more sophisticated discriminative phrase training method for the WMT 2014 German→English task. Similar to (He and Deng, 2012), 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. To that end, the training data is decoded to generate 100-best lists. We apply a leave-one-out heuristic (Wuebker et al., 2010) 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 run MERT, evaluate on the development set and select the best performing iteration. In this work, we perform two rounds of discriminative training on two separate data sets. In the first round, training is performed on the concatenation of newstest2008 through newstest2010 and an automatic selection from the News-commentary, Europarl and Common Crawl corpora. The selection is based on cross-entropy difference of language models and IBM-1 models as described by Mansour et al. (2011) and contains 258K sentence pairs. The training took 4.5 hours for 30 iterations. On top of the final phrase-based systems, a second round of discriminative training is run on the full news-commentary corpus concatenated with newstest2008 through newstest2010.

3.5 A Phrase Orientation Model for Hierarchical Machine Translation

In Huck et al. (2013b) a lexicalized reordering model for hierarchical 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). Since improvements were reported on a Chinese→English translation task, we investigate the impact of this model on a European language pair. As in German the word order is more flexible compared with the target language English, we expect that an additional reordering model could improve the translation quality. In our experiments we use the same settings which worked best in (Huck et al., 2013b).

4 Setup

We trained the phrase-based and the hierarchical translation system on all available bilingual training data. Corpus statistics can be found in the last row of Table 2. The language model are 4-grams trained on the respective target side of the bilingual data, $\frac{1}{2}$ of the Shuffled News Crawl corpus, $\frac{1}{4}$ of the 10^9 French-English corpus and $\frac{1}{2}$ of the LDC Gigaword Fifth Edition corpus. The monolingual data selection is based on cross-entropy difference as described in (Moore and Lewis, 2010). For the baseline language model, we trained separate models for each corpus, which were then interpolated. For our final experiments, we also trained a single unpruned language model on the concatenation of all monolingual data with KenLM.

Table 3: Results (truecase) for the German→English translation task. BLEU and TER are given in percentage. All HPBT setups are tuned on the concatenation of newstest2012 and newstest2013. The very first SCSS setups are optimized on newstest2012 only.

	newstest2011		newstest2012		newstest2013	
	BLEU	TER	BLEU	TER	BLEU	TER
SCSS +HRM	22.4	60.1	23.7	59.0	25.9	55.7
+wcLM	22.8	59.6	24.0	58.6	26.3	55.4
+1st round discr.	23.0	59.5	24.2	58.2	26.8	55.1
+tune11+12.	23.4	59.5	24.2	58.6	26.8	55.2
+unprunedLM	23.6	59.5	24.2	58.6	27.1	55.0
+2nd round discr.	23.7	59.5	24.4	58.5	27.2	55.0
HPBT baseline	23.3	59.9	24.2	58.9	26.7	55.6
+wcLM	23.4	59.8	24.1	58.9	26.8	55.6
+HRM	23.3	60.0	24.2	58.9	26.9	55.5
+HRM +wcLM	23.3	59.9	24.1	59.1	26.7	55.9

4.1 Experimental Results

The results of the phrase-based system (SCSS) as well as the hierarchical phrase-based system (HPBT) are summarized in Table 3.

The phrase-based baseline system, which includes the hierarchical reordering model by (Galley and Manning, 2008) and is tuned on newstest2012, reaches a performance of 25.9% BLEU on newstest2013. Adding the word class language model improves performance by 0.4% BLEU absolute and the first round of discriminative phrase training by 0.5% BLEU absolute. Next, we switched to tuning on a concatenation of newstest2011 and newstest2012, which we expect to be more reliable with respect to unseen data. Although the BLEU score does not improve and TER goes up slightly, we kept this tuning set in the subsequent setups, as it yielded longer translations, which in our experience will usually be preferred by human evaluators. Switching from the interpolated language model to the unpruned language model trained with KenLM on the full concatenated monolingual training data in a single pass gained us another 0.3% BLEU. For the final system, we ran a second round of discriminative training on different training data (cf. Section 3.4), which increased performance by 0.1% BLEU to the final score 27.2.

For the phrase-based system, we also experimented with weighted phrase extraction (Mansour and Ney, 2012), but did not observe improvements.

The hierarchical phrase-based baseline without any additional model is on the same level as the phrase-based system including the word class language model, hierarchical reordering model and discriminative phrase training in terms of BLEU. However, extending the system with a word class language model or the additional reordering models does not seem to help. Even the combination of both models does not improve the translation quality. Note, that the hierarchical system was tuned on the concatenation newstest2011 and newstest2012. The final system employs both word class language model and hierarchical reordering model.

Both phrase-based and hierarchical phrase-based final systems are used in the EU-Bridge system combination (Freitag et al., 2014).

5 Conclusion

For the participation in the WMT 2014 shared translation task, RWTH experimented with both phrase-based and hierarchical translation systems. For both approaches, we applied a hierarchical phrase reordering model and a word class language model. For the phrase-based system we employed discriminative phrase training. Additionally, improvements of our preprocessing pipeline compared to our WMT 2013 setup were described. New introduced categories lead to a lower amount of out-of-vocabulary words. Filtering the corpus for wrong languages gives us lower vocabulary sizes for source and target without loosing any performance.

Acknowledgments

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

Furthermore, this material is partially based upon work supported by the DARPA BOLT project under Contract No. HR0011- 12-C-0015. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

Stanley F. Chen and Joshua Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, Massachusetts, USA, August.

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

Markus Freitag, Stephan Peitz, Joern Wuebker, Hermann Ney, Matthias Huck, Rico Sennrich, Nadir Durrani, Maria Nadejde, Philip Williams, Philipp Koehn, Teresa Herrmann, Eunah Cho, and Alex Waibel. 2014. EU-BRIDGE MT: Combined Machine Translation. In *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, June.

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.

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.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

Matthias Huck, Jan-Thorsten Peter, Markus Freitag, Stephan Peitz, and Hermann Ney. 2012. Hierarchical Phrase-Based Translation with Jane 2. *The Prague Bulletin of Mathematical Linguistics*, 98:37–50, October.

Matthias Huck, David Vilar, Markus Freitag, and Hermann Ney. 2013a. A Performance Study of Cube Pruning for Large-Scale Hierarchical Machine Translation. In *Proceedings of the NAACL 7th Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 29–38, Atlanta, Georgia, USA, June.

Matthias Huck, Joern Wuebker, Felix Rietig, and Hermann Ney. 2013b. A phrase orientation model for hierarchical machine translation. In *ACL 2013 Eighth Workshop on Statistical Machine Translation*, pages 452–463, Sofia, Bulgaria, August.

Reinhard Kneser and Hermann Ney. 1995. Improved Backing-Off for M-gram Language Modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, May.

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

Saab Mansour and Hermann Ney. 2012. A Simple and Effective Weighted Phrase Extraction for Machine Translation Adaptation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 193–200, Hong Kong, December.

Saab Mansour, Joern Wuebker, and Hermann Ney. 2011. Combining Translation and Language Model Scoring for Domain-Specific Data Filtering. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 222–229, San Francisco, California, USA, December.

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. 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.

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.

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.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, Colorado, 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. 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.

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.