The RWTH Aachen Machine Translation System for WMT 2010

Carmen Heger, Joern Wuebker, Matthias Huck, Gregor Leusch,
Saab Mansour, Daniel Stein and Hermann Ney
RWTH Aachen University
Aachen, Germany
surname@cs.rwth-aachen.de

Abstract

In this paper we describe the statistical machine translation system of the RWTH Aachen University developed for the translation task of the Fifth Workshop on Statistical Machine Translation. State-of-the-art phrase-based and hierarchical statistical MT systems are augmented with appropriate morpho-syntactic enhancements, as well as alternative phrase training methods and extended lexicon models. For some tasks, a system combination of the best systems was used to generate a final hypothesis. We participated in the constrained condition of German-English and French-English in each translation direction.

1 Introduction

This paper describes the statistical MT system used for our participation in the WMT 2010 shared translation task. We used it as an opportunity to incorporate novel methods which have been investigated at RWTH over the last year and which have proven to be successful in other evaluations.

For all tasks we used standard alignment and training tools as well as our in-house phrase-based and hierarchical statistical MT decoders. When German was involved, morpho-syntactic preprocessing was applied. An alternative phrase-training method and additional models were tested and investigated with respect to their effect for the different language pairs. For two of the language pairs we could improve performance by system combination.

An overview of the systems and models will follow in Section 2 and 3, which describe the baseline architecture, followed by descriptions of the additional system components. Morpho-syntactic analysis and other preprocessing issues are covered by Section 4. Finally, translation results for the different languages and system variants are presented in Section 5.

2 Translation Systems

For the WMT 2010 Evaluation we used standard phrase-based and hierarchical translation systems. Alignments were trained with a variant of GIZA++. Target language models are 4-gram language models trained with the SRI toolkit, using Kneser-Ney discounting with interpolation.

2.1 Phrase-Based System

Our phrase-based translation system is similar to the one described in (Zens and Ney, 2008). Phrase pairs are extracted from a word-aligned bilingual corpus and their translation probability in both directions is estimated by relative frequencies. Additional models include a standard n-gram language model, phrase-level IBM1, word-, phrase- and distortion-penalties and a discriminative reordering model as described in (Zens and Ney, 2006).

2.2 Hierarchical System

Our hierarchical phrase-based system is similar to the one described in (Chiang, 2007). It allows for gaps in the phrases by employing a context-free grammar and a CYK-like parsing during the decoding step. It has similar features as the phrase-based system mentioned above. For some systems, we only allowed the non-terminals in hierarchical phrases to be substituted with initial phrases as in (Iglesias et al., 2009), which gave better results on some language pairs. We will refer to this as “shallow rules”.

2.3 System Combination

The RWTH approach to MT system combination of the French→English systems as well as the German→English systems is a refined version of the ROVER approach in ASR (Fiscus, 1997) with

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 99–103,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

	German→English		French→English		English→French	
	BLEU	# Phrases	BLEU	# Phrases	BLEU	# Phrases
Standard	19.7	128M	25.5	225M	23.7	261M
FA	20.0	12M	25.9	35M	24.0	33M

Table 1: BLEU scores on Test and phrase table sizes with and without forced alignment (FA). For German→English and English→French phrase table interpolation was applied.

additional steps to cope with reordering between different hypotheses, and to use true casing information from the input hypotheses. The basic concept of the approach has been described by Matusov et al. (2006). Several improvements have been added later (Matusov et al., 2008). This approach includes an enhanced alignment and reordering framework. Alignments between the systems are learned by GIZA++, a one-to-one alignment is generated from the learned state occupation probabilities.

From these alignments, a confusion network (CN) is then built using one of the hypotheses as "skeleton" or "primary" hypothesis. We do not make a hard decision on which of the hypotheses to use for that, but instead combine all possible CNs into a single lattice. Majority voting on the generated lattice is performed using the prior probabilities for each system as well as other statistical models such as a special trigram language model. This language model is also learned on the input hypotheses. The intention is to favor longer phrases contained in individual hypotheses. The translation with the best total score within this lattice is selected as consensus translation. Scaling factors of these models are optimized similar to MERT using the Downhill Simplex algorithm. As the objective function for this optimization, we selected a linear combination of BLEU and TER with a weight of 2 on the former; a combination that has proven to deliver stable results on several MT evaluation measures in preceding experiments.

In contrast to previous years, we now include a separate consensus true casing step to exploit the true casing capabilities of some of the input systems: After generating a (lower cased) consensus translation from the CN, we sum up the counts of different casing variants of each word in a sentence over the input hypotheses, and use the majority casing over those. In previous experiments, this showed to work significantly better than using a fixed non-consensus true caser, and maintains flexibility on the input systems.

3 New Additional Models

3.1 Forced Alignment

For the German→English, French→English and English→French language tasks we applied a forced alignment procedure to train the phrase translation model with the EM algorithm, similar to the one described in (DeNero et al., 2006). Here, the phrase translation probabilities are estimated from their relative frequencies in the phrase-aligned training data. The phrase alignment is produced by a modified version of the translation decoder. In addition to providing a statistically well-founded phrase model, this has the benefit of producing smaller phrase tables and thus allowing more rapid experiments. For the language pairs German→English and English→French the best results were achieved by log-linear interpolation of the standard phrase table with the generative model. For French→English we directly used the model trained by forced alignment. A detailed description of the training procedure is given in (Wuebker et al., 2010). Table 1 shows the system performances and phrase table sizes with the standard phrase table and the one trained with forced alignment after the first EM iteration. We can see that the generative model reduces the phrase table size by 85-90% while increasing performance by 0.3% to 0.4% BLEU.

3.2 Extended Lexicon Models

In previous work, RWTH was able to show the positive impact of extended lexicon models that cope with lexical context beyond the limited horizon of phrase pairs and n-gram language models.

Mauser et al. (2009) report improvements of up to +1% in BLEU on large-scale systems for Chinese→English and Arabic→English by incorporating discriminative and trigger-based lexicon models into a state-of-the-art phrase-based decoder. They discuss how the two types of lexicon models help to select content words by capturing long-distance effects.

The triplet model is a straightforward extension of the IBM model 1 with a second trigger, and like the former is trained iteratively using the EM algorithm. In search, the triggers are usually on the source side, i.e., $p(e|f,f')$ is modeled. The path-constrained triplet model restricts the first source trigger to the aligned target word, whereas the second trigger can move along the whole source sentence. See (Hasan et al., 2008) for a detailed description and variants of the model and its training.

For the WMT 2010 evaluation, triplets modeling $p(e|f,f')$ were trained and applied directly in search for all relevant language pairs. Path-constrained models were trained on the in-domain news-commentary data only and on the news-commentary plus the Europarl data. Although experience from similar setups indicates that triplet lexicon models can be beneficial for machine translation between the languages English, French, and German, on this year's WMT translation tasks slight improvements on the development sets did not or only partially carry over to the held-out test sets. Nevertheless, systems with triplets were used for system combination, as extended lexicon models often help to predict content words and to capture long-range dependencies. Thus they can help to find a strong consensus hypothesis.

3.3 Unsupervised Training

Due to the small size of the English→German resources available for language modeling as well as for lexicon extraction, we decided to apply the unsupervised adaptation suggested in (Schwenk and Senellart, 2009). We use a baseline SMT system to translate in-domain monolingual source data, filter the translations according to a decoder score normalized by sentence length, add this synthetic bilingual data to the original one and rebuild the SMT system from scratch.

The motivation behind the method is that the phrase table will adapt to the genre, and thus let phrases which are domain related have higher probabilities. Two phenomena are observed from phrase tables and the corresponding translations:

- Phrase translation probabilities are changed, making the system choose better phrase translation candidates.
- Phrases which appear repeatedly in the domain get higher probabilities, so that the decoder can better segment the sentence.

	Running Words	
	English	German
Bilingual	44.3M	43.4M
Dict.	1.4M	1.2M
AFP	610.7M	
AFP unsup.	152.0M	157.3M

Table 2: Overview on data for unsupervised training.

	BLEU	
	Dev	Test
baseline	15.0	14.7
+dict.	15.1	14.6
+unsup.+dict	15.4	14.9

Table 3: Results for unsupervised training method.

To implement this idea, we translate the AFP part of the English LDC Gigaword v4.0 and obtain the synthetic data.

To decrease the number of OOV words, we use dictionaries from the stardict directory as additional bilingual data to translate the AFP corpus. We filter sentences with OOV words and sentences longer than 100 tokens. A summary of the additional data used is shown in Table 2.

We tried to use the best 10%, 20% and 40% of the synthetic data, where the 40% option worked best. A summary of the results is given in Table 3.

Although this is our best result for the English→German task, it was not submitted, because the use of the dictionary is not allowed in the constrained track.

4 Preprocessing

4.1 Large Parallel Data

In addition to the provided parallel Europarl and news-commentary corpora, also the large French-English news corpus (about 22.5 Mio. sentence pairs) and the French-English UN corpus (about 7.2 Mio. sentence pairs) were available. Since model training and tuning with such large corpora takes a very long time, we extracted about 2 Mio. sentence pairs of both of these corpora. We filter sentences with the following properties:

- Only sentences of minimum length of 4 tokens were considered.
- At least 92% of the vocabulary of each sentence occur in the development set.
- The ratio of the vocabulary size of a sentence and the number of its tokens is minimum 80%.

4.2 Morpho-Syntactic Analysis

German, as a flexible and morphologically rich language, raises a couple of problems in machine translation. We picked two major problems and tackled them with morpho-syntactic pre- and post-processing: compound splitting and long-range verb reordering.

For the translation from German into English, German compound words were split using the frequency-based method described in (Koehn and Knight, 2003). Thereby, we forbid certain words and syllables to be split. For the other translation direction, the English text was first translated into the modified German language with split compounds. The generated output was then postprocessed by re-merging the previously generated components using the method described in (Popović et al., 2006).

Additionally, for the German→English phrase-based system, the long-range POS-based reordering rules described in (Popović and Ney, 2006) were applied on the training and test corpora as a preprocessing step. Thereby, German verbs which occur at the end of a clause, like infinitives and past participles, are moved towards the beginning of that clause. With this, we improved our baseline phrase-based system by 0.6% BLEU.

5 Experimental Results

For all translation directions, we used the provided parallel corpora (Europarl, news) to train the translation models and the monolingual corpora to train the language models. We improved the French-English systems by enriching the data with parts of the large addional data, extracted with the method described in Section 4.1. Depending on the system this gave an improvement of 0.2-0.7% BLEU. We also made use of the large giga-news as well as the LDC Gigaword corpora for the French and English language models. All systems were optimized for BLEU score on the development data, `newstest2008`. The `newstest2009` data is used as a blind test set.

In the following, we will give the BLEU scores for all language tasks of the baseline system and the best setup for both, the phrase-based and the hierarchical system. We will use the following notations to indicate the several methods we used:

(+POS)	POS-based verb reordering
(+mero)	maximum entropy reordering
(+giga)	including giga-news and LDC Gigaword in LM
(fa)	trained by forced alignment
(shallow)	allow only shallow rules

We applied system combination of up to 6 systems with several setups. The submitted systems are marked in tables 4-7.

6 Conclusion

For the participation in the WMT 2010 shared translation task, RWTH used state-of-the-art phrase-based and hierarchical translation systems. To deal with the rich morphology and word order differences in German, compound splitting and long range verb reordering were applied in a preprocessing step. For the French-English language pairs, RWTH extracted parts of the large news corpus and the UN corpus as additional training data. Further, training the phrase translation model with forced alignment yielded improvements in BLEU. To obtain the final hypothesis for the French→English and German→English

	BLEU	
	Dev	Test
phrase-based baseline	19.9	19.2
phrase-based (+POS+mero+giga)	21.0	20.3
hierarchical baseline	20.2	19.6
hierarchical (+giga)	20.5	20.1
system combination	21.4	20.4

Table 4: Results for the German→English task.

	BLEU	
	Dev	Test
phrase-based baseline	14.8	14.5
phrase-based (+mero)	15.0	14.7
hierarchical baseline	14.2	13.9
hierarchical (shallow)	14.5	14.3

Table 5: Results for the English→German task.

	BLEU	
	Dev	Test
phrase-based baseline	21.8	25.1
phrase-based (fa+giga)	23.0	26.1
hierarchical baseline	21.9	25.0
hierarchical (shallow+giga)	22.7	25.6
system combination	23.1	26.1

Table 6: Results for the French→English task.

	BLEU	
	Dev	Test
phrase-based baseline	20.9	23.2
phrase-based (fa+mero+giga)	23.0	24.6
hierarchical baseline	20.6	22.5
hierarchical (shallow,+giga)	22.4	24.3

Table 7: Results for the English→French task.

language pairs, RWTH applied system combination. Altogether, by application of these methods RWTH was able to increase performance in BLEU by 0.8% for German→English, 0.2% for English→German, 1.0% for French→English and 1.4% for English→French on the test set over the respective baseline systems.

Acknowledgments

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

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

J. DeNero, D. Gillick, J. Zhang, and D. Klein. 2006. Why Generative Phrase Models Underperform Surface Heuristics. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 31–38.

J.G. Fiscus. 1997. A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding*.

S. Hasan, J. Ganitkevitch, H. Ney, and J. Andrés-Ferrer. 2008. Triplet Lexicon Models for Statistical Machine Translation. In *Proceedings of Emperical Methods of Natural Language Processing*, pages 372–381.

G. Iglesias, A. de Gispert, E.R. Banga, and W. Byrne. 2009. Rule Filtering by Pattern for Efficient Hierarchical Translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 380–388.

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

E. Matusov, N. Ueffing, and H. Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 33–40.

E. Matusov, G. Leusch, R.E. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y.-S. Lee, J.B. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney. 2008. System Combination for Machine Translation of Spoken and Written Language. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1222–1237.

A. Mauser, S. Hasan, and H. Ney. 2009. Extending Statistical Machine Translation with Discriminative and Trigger-Based Lexicon Models. In *Conference on Empirical Methods in Natural Language Processing*, pages 210–217.

M. Popović and H. Ney. 2006. POS-based Word Reorderings for Statistical Machine Translation. In *International Conference on Language Resources and Evaluation*, pages 1278–1283.

M. Popović, D. Stein, and H. Ney. 2006. Statistical Machine Translation of German Compound Words. In *FinTAL - 5th International Conference on Natural Language Processing, Springer Verlag, LNCS*, pages 616–624.

H. Schwenk and J. Senellart. 2009. Translation Model Adaptation for an Arabic/French News Translation System by Lightly-Supervised Training. In *MT Summit XII*.

J. Wuebker, A. Mauser, and H. Ney. 2010. Training Phrase Translation Models with Leaving-One-Out. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. To appear.

R. Zens and H. Ney. 2006. Discriminative Reordering Models for Statistical Machine Translation. In *Workshop on Statistical Machine Translation*, pages 55–63.

R. Zens and H. Ney. 2008. Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation. In *International Workshop on Spoken Language Translation*.