The Karlsruhe Institute of Technology Translation Systems for the WMT 2014

Teresa Herrmann, Mohammed Mediani, Eunah Cho, Thanh-Le Ha,
Jan Niehues, Isabel Slawik, Yuqi Zhang and Alex Waibel
Institute for Anthropomatics
KIT - Karlsruhe Institute of Technology
firstname.lastname@kit.edu

Abstract

In this paper, we present the KIT systems participating in the Shared Translation Task translating between English↔German and English↔French. All translations are generated using phrase-based translation systems, using different kinds of word-based, part-of-speech-based and cluster-based language models trained on the provided data. Additional models further include bilingual language models, reordering models based on part-of-speech tags and syntactic trees, as well as a lexicalized reordering model. In order to make use of noisy web-crawled data, we apply filtering and data selection methods for language modelling. A discriminative word lexicon using source context information proved beneficial for all translation directions.

1 Introduction

We describe the KIT systems for the Shared Translation Task of the ACL 2014 Ninth Workshop on Statistical Machine Translation. We participate in the English↔German and English↔French translation directions, using a phrase-based decoder with lattice input.

The paper is organized as follows: the next section describes the data used for each translation direction. Section 3 gives a detailed description of our systems including all the models. The translation results for all directions are presented afterwards and we close with a conclusion.

2 Data

We utilize the provided EPPS, NC and Common Crawl parallel corpora for English→German and German→English, plus Giga for English→French and French→English. The monolingual part of those parallel corpora, the News Shuffle corpus for all four directions and additionally the Gigaword corpus for English→French and German→English are used as monolingual training data for the different language models. For optimizing the system parameters, newstest2012 and newstest2013 are used as development and test data respectively.

3 System Description

Before training we perform a common preprocessing of the raw data, which includes removing long sentences and sentences with a length mismatch exceeding a certain threshold. Afterwards, we normalize special symbols, dates, and numbers. Then we perform smart-casing of the first letter of every sentence. Compound splitting (Koehn and Knight, 2003) is performed on the source side of the corpus for German→English translation. In order to improve the quality of the web-crawled Common Crawl corpus, we filter out noisy sentence pairs using an SVM classifier for all four translation tasks as described in Mediani et al. (2011).

Unless stated otherwise, we use 4-gram language models (LM) with modified Kneser-Ney smoothing, trained with the SRILM toolkit (Stolcke, 2002). All translations are generated by an in-house phrase-based translation system (Vogel, 2003), and we use Minimum Error Rate Training (MERT) as described in Venugopal et al. (2005) for optimization. The word alignment of the parallel corpora is generated using the GIZA++ Toolkit (Och and Ney, 2003) for both directions. Afterwards, the alignments are combined using the grow-diag-final-and heuristic. For English→German, we use discriminative word alignment as described in Niehues and Vogel (2008). The phrase table (PT) is built using the Moses toolkit (Koehn et al., 2007). The scoring for the small data sets (German↔English) is also done by the Moses toolkit, whereas the bigger

sets (French↔English) are scored by our in-house parallel phrase scorer (Mediani et al., 2012a). The phrase pair probabilities are computed using modified Kneser-Ney smoothing as described in Foster et al. (2006).

Since German is a highly inflected language, we try to alleviate the out-of-vocabulary problem through quasi-morphological operations that change the lexical entry of a known word form to an unknown word form as described in Niehues and Waibel (2011).

3.1 Word Reordering Models

We apply automatically learned reordering rules based on part-of-speech (POS) sequences and syntactic parse tree constituents to perform source sentence reordering according to the target language word order. The rules are learned from a parallel corpus with POS tags (Schmid, 1994) for the source side and a word alignment to learn reordering rules that cover short range (Rottmann and Vogel, 2007) and long range reorderings (Niehues and Kolss, 2009). In addition, we apply a tree-based reordering model (Herrmann et al., 2013) to better address the differences in word order between German and English. Here, a word alignment and syntactic parse trees (Rafferty and Manning, 2008; Klein and Manning, 2003) for the source side of the training corpus are required to learn rules on how to reorder the constituents in the source sentence. The POS-based and tree-based reordering rules are applied to each input sentence before translation. The resulting reordered sentence variants as well as the original sentence are encoded in a reordering lattice. The lattice, which also includes the original position of each word, is used as input to the decoder.

In order to acquire phrase pairs matching the reordered sentence variants, we perform lattice phrase extraction (LPE) on the training corpus where phrase are extracted from the reordered word lattices instead of the original sentences.

In addition, we use a lexicalized reordering model (Koehn et al., 2005) which stores reordering probabilities for each phrase pair. During decoding the lexicalized reordering model determines the reordering orientation of each phrase pair at the phrase boundaries. The probability for the respective orientation with respect to the original position of the words is included as an additional score in the log-linear model of the translation system.

3.2 Adaptation

In the French→English and English→French systems, we perform adaptation for translation models as well as for language models. The EPPS and NC corpora are used as in-domain data for the direction English→French, while NC corpus is the in-domain data for French→English.

Two phrase tables are built: one is the out-of-domain phrase table, which is trained on all corpora; the other is the in-domain phrase table, which is trained on in-domain data. We adapt the translation model by using the scores from the two phrase tables with the backoff approach described in Niehues and Waibel (2012). This results in a phrase table with six scores, the four scores from the general phrase table as well as the two conditional probabilities from the in-domain phrase table. In addition, we take the union of the candidate phrase pairs collected from both phrase tables A detailed description of the union method can be found in Mediani et al. (2012b).

The language model is adapted by log-linearly combining the general language model and an in-domain language model. We train a separate language model using only the in-domain data. Then it is used as an additional language model during decoding. Optimal weights are set during tuning by MERT.

3.3 Special Language Models

In addition to word-based language models, we use different types of non-word language models for each of the systems. With the help of a bilingual language model (Niehues et al., 2011) we are able to increase the bilingual context between source and target words beyond phrase boundaries. This language model is trained on bilingual tokens created from a target word and all its aligned source words. The tokens are ordered according to the target language word order.

Furthermore, we use language models based on fine-grained part-of-speech tags (Schmid and Laws, 2008) as well as word classes to alleviate the sparsity problem for surface words. The word classes are automatically learned by clustering the words of the corpus using the MKCLS algorithm (Och, 1999). These n-gram language models are trained on the target language corpus, where the words have been replaced either by their

corresponding POS tag or cluster ID. During decoding, these language models are used as additional models in the log-linear combination.

The data selection language model is trained on data automatically selected using cross-entropy differences between development sets from previous WMT workshops and the noisy crawled data (Moore and Lewis, 2010). We selected the top 10M sentences to train this language model.

3.4 Discriminative Word Lexicon

A discriminative word lexicon (DWL) models the probability of a target word appearing in the translation given the words of the source sentence. DWLs were first introduced by Mauser et al. (2009). For every target word, they train a maximum entropy model to determine whether this target word should be in the translated sentence or not using one feature per source word.

We use two simplifications of this model that have shown beneficial to translation quality and training time in the past (Mediani et al., 2011). Firstly, we calculate the score for every phrase pair before translating. Secondly, we restrict the negative training examples to words that occur within matching phrase pairs.

In this evaluation, we extended the DWL with n-gram source context features proposed by Niehues and Waibel (2013). Instead of representing the source sentence as a bag-of-words, we model it as a bag-of-n-grams. This allows us to include information about source word order in the model. We used one feature per n-gram up to the order of three and applied count filtering for bigrams and trigrams.

4 Results

This section presents the participating systems used for the submissions in the four translation directions of the evaluation. We describe the individual components that form part of each of the systems and report the translation qualities achieved during system development. The scores are reported in case-sensitive BLEU (Papineni et al., 2002).

4.1 English-French

The development of our English→French system is shown in Table 1.

It is noteworthy that, for this direction, we chose to tune on a subset of 1,000 pairs from newstest2012, due to the long time the whole set takes to be decoded. In a preliminary set of experiments (not reported here), we found no significant differences between tuning on the small or the big development sets. The translation model of the baseline system is trained on the whole parallel data after filtering (EPPS, NC, Common Crawl, Giga). The same data was also used for language modeling. We also use POS-based reordering.

The biggest improvement was due to using two additional language models. One consists of a log-linear interpolation of individual language models trained on the target side of the parallel data, the News shuffle, Gigaword and NC corpora. In addition, an in-domain language model trained only on NC data is used. This improves the score by more than 1.4 points. Adaptation of the translation model towards a smaller model trained on EPPS and NC brings an additional 0.3 points.

Another 0.3 BLEU points could be gained by using other special language models: a bilingual language model together with a 4-gram cluster language model (trained on all monolingual data using the MKCLS tool and 500 clusters). Incorporating a lexicalized reordering model into the system had a very noticeable effect on test namely more than half a BLEU point.

Finally, using a discriminative word lexicon with source context has a very small positive effect on the test score, however more than 0.3 on dev. This final configuration was the basis of our submitted official translation.

System	Dev	Test
Baseline	15.63	27.61
+ Big LMs	16.56	29.02
+ PT Adaptation	16.77	29.32
+ Bilingual + Cluster LM	16.87	29.64
+ Lexicalized Reordering	16.92	30.17
+ Source DWL	17.28	**30.19**

Table 1: Experiments for English→French

4.2 French-English

Several experiments were conducted for the French→English translation system. They are summarized in Table 2.

The baseline system is essentially a phrase-based translation system with some preprocessing steps on the source side and utilizing the short-range POS-based reordering on all parallel

data and fine-grained monolingual corpora such as EPPS and NC.

Adapting the translation model using a small in-domain phrase table trained on NC data only helps us gain more than 0.4 BLEU points.

Using non-word language models including a bilingual language model and a 4-gram 50-cluster language model trained on the whole parallel data attains 0.24 BLEU points on the test set.

Lexicalized reordering improves our system on the development set by 0.3 BLEU points but has less effect on the test set with a minor improvement of around 0.1 BLEU points.

We achieve our best system, which is used for the evaluation, by adding a DWL with source context yielding 31.54 BLEU points on the test set.

System	Dev	Test
Baseline	30.16	30.70
+ LM Adaptation	30.58	30.94
+ PT Adaptation	30.69	31.14
+ Bilingual + Cluster LM	30.85	31.38
+ Lexicalized Reordering	31.14	31.46
+ Source DWL	31.19	**31.54**

Table 2: Experiments for French→English

4.3 English-German

Table 3 presents how the English-German translation system is improved step by step.

In the baseline system, we used parallel data which consists of the EPPS and NC corpora. The phrase table is built using discriminative word alignment. For word reordering, we use word lattices with long range reordering rules. Five language models are used in the baseline system; two word-based language models, a bilingual language model, and two 9-gram POS-based language models. The two word-based language models use 4-gram context and are trained on the parallel data and the filtered Common Crawl data separately, while the bilingual language model is built only on the Common Crawl corpus. The two POS-based language models are also based on the parallel data and the filtered crawled data, respectively.

When using a 9-gram cluster language model, we get a slight improvement. The cluster is trained with 1,000 classes using EPPS, NC, and Common Crawl data.

We use the filtered crawled data in addition to the parallel data in order to build the phrase table; this gave us 1 BLEU point of improvement.

The system is improved by 0.1 BLEU points when we use lattice phrase extraction along with lexicalized reordering rules.

Tree-based reordering rules improved the system performance further by another 0.1 BLEU points.

By reducing the context of the two POS-based language models from 9-grams to 5-grams and shortening the context of the language model trained on word classes to 4-grams, the score on the development set hardly changes but we can see a slightly improvement for the test case.

Finally, we use the DWL with source context and build a big bilingual language model using both the crawled and parallel data. By doing so, we improved the translation performance by another 0.3 BLEU points. This system was used for the translation of the official test set.

System	Dev	Test
Baseline	16.64	18.60
+ Cluster LM	16.76	18.66
+ Common Crawl Data	17.27	19.66
+ LPE + Lexicalized Reordering	17.45	19.75
+ Tree Rules	17.53	19.85
+ Shorter n-grams	17.55	19.92
+ Source DWL + Big BiLM	17.82	**20.21**

Table 3: Experiments for English→German

4.4 German-English

Table 4 shows the development steps of the German-English translation system.

For the baseline system, the training data of the translation model consists of EPPS, NC and the filtered parallel crawled data. The phrase table is built using GIZA++ word alignment and lattice phrase extraction. All language models are trained with SRILM and scored in the decoding process with KenLM (Heafield, 2011). We use word lattices generated by short and long range reordering rules as input to the decoder. In addition, a bilingual language model and a target language model trained on word clusters with 1,000 classes are included in the system.

Enhancing the word reordering with tree-based reordering rules and a lexicalized reordering model improved the system performance by 0.6 BLEU points.

Adding a language model trained on selected data from the monolingual corpora gave another small improvement.

The DWL with source context increased the score on the test set by another 0.5 BLEU points and applying morphological operations to unknown words reduced the out-of-vocabulary rate, even though no improvement in BLEU can be observed. This system was used to generate the translation submitted to the evaluation.

System	Dev	Test
Baseline	24.40	26.34
+ Tree Rules	24.71	26.86
+ Lexicalized Reordering	24.89	26.93
+ LM Data Selection	24.96	27.03
+ Source DWL	25.32	27.53
+ Morphological Operations	-	**27.53**

Table 4: Experiments for German→English

5 Conclusion

In this paper, we have described the systems developed for our participation in the Shared Translation Task of the WMT 2014 evaluation for English↔German and English↔French. All translations were generated using a phrase-based translation system which was extended by additional models such as bilingual and fine-grained part-of-speech language models. Discriminative word lexica with source context proved beneficial in all four language directions.

For English-French translation using a smaller development set performed reasonably well and reduced development time. The most noticeable gain comes from log-linear interpolation of multiple language models.

Due to the large amounts and diversity of the data available for French-English, adaptation methods and non-word language models contribute the major improvements to the system.

For English-German translation, the crawled data and a DWL using source context to guide word choice brought most of the improvements.

Enhanced word reordering models, namely tree-based reordering rules and a lexicalized reordering model as well as the source-side features for the discriminative word lexicon helped improve the system performance for German-English translation.

In average we achieved an improvement of over 1.5 BLEU over the respective baselines for all our systems.

Acknowledgments

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

References

George F. Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *Proceedings of the 2006 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, Sydney, Australia.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, United Kingdom.

Teresa Herrmann, Jan Niehues, and Alex Waibel. 2013. Combining Word Reordering Methods on different Linguistic Abstraction Levels for Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, Altanta, Georgia, USA.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary.

Philipp Koehn, Amittai Axelrod, Alexandra B. Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *Proceedings of the Second International Workshop on Spoken Language Translation (IWSLT 2005)*, Pittsburgh, PA, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Demonstration Session*, Prague, Czech Republic.

Arne Mauser, Saša Hasan, and Hermann Ney. 2009. Extending Statistical Machine Translation with Discriminative and Trigger-based Lexicon Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Suntec, Singapore.

Mohammed Mediani, Eunah Cho, Jan Niehues, Teresa Herrmann, and Alex Waibel. 2011. The KIT English-French Translation systems for IWSLT 2011. In *Proceedings of the Eight International Workshop on Spoken Language Translation (IWSLT 2011)*, San Francisco, CA, USA.

Mohammed Mediani, Jan Niehues, and Alex Waibel. 2012a. Parallel Phrase Scoring for Extra-large Corpora. In *The Prague Bulletin of Mathematical Linguistics*, number 98.

Mohammed Mediani, Yuqi Zhang, Thanh-Le Ha, Jan Niehues, Eunach Cho, Teresa Herrmann, Rainer Kärgel, and Alexander Waibel. 2012b. The KIT Translation Systems for IWSLT 2012. In *Proceedings of the Ninth International Workshop on Spoken Language Translation (IWSLT 2012)*, Hong Kong, HK.

R.C. Moore and W. Lewis. 2010. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, Uppsala, Sweden.

Jan Niehues and Muntsin Kolss. 2009. A POS-Based Model for Long-Range Reorderings in SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT 2009)*, Athens, Greece.

Jan Niehues and Stephan Vogel. 2008. Discriminative Word Alignment via Alignment Matrix Modeling. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT 2008)*, Columbus, OH, USA.

Jan Niehues and Alex Waibel. 2011. Using Wikipedia to Translate Domain-specific Terms in SMT. In *Proceedings of the Eight International Workshop on Spoken Language Translation (IWSLT 2008)*, San Francisco, CA, USA.

J. Niehues and A. Waibel. 2012. Detailed Analysis of Different Strategies for Phrase Table Adaptation in SMT. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA 2012)*, San Diego, CA, USA.

J. Niehues and A. Waibel. 2013. An MT Error-Driven Discriminative Word Lexicon using Sentence Structure Features. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria.

Jan Niehues, Teresa Herrmann, Stephan Vogel, and Alex Waibel. 2011. Wider Context by Using Bilingual Language Models in Machine Translation. In *Sixth Workshop on Statistical Machine Translation (WMT 2011)*, Edinburgh, Scotland, United Kingdom.

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

Franz Josef Och. 1999. An Efficient Method for Determining Bilingual Word Classes. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL 1999)*, Bergen, Norway.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176 (W0109-022), IBM Research Division, T. J. Watson Research Center.

Anna N. Rafferty and Christopher D. Manning. 2008. Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In *Proceedings of the Workshop on Parsing German*, Columbus, OH, USA.

Kay Rottmann and Stephan Vogel. 2007. Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, Skövde, Sweden.

Helmut Schmid and Florian Laws. 2008. Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. In *International Conference on Computational Linguistics (COLING 2008)*, Manchester, Great Britain.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, Manchester, United Kingdom.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *International Conference on Spoken Language Processing*, Denver, Colorado, USA.

Ashish Venugopal, Andreas Zollman, and Alex Waibel. 2005. Training and Evaluation Error Minimization Rules for Statistical Machine Translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, Michigan, USA.

Stephan Vogel. 2003. SMT Decoder Dissected: Word Reordering. In *International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China.