The Karlsruhe Institute of Technology Translation Systems for the WMT 2015

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

Abstract

In this paper, the KIT systems submitted to the Shared Translation Task are presented. We participated in two translation directions: from German to English and from English to German. Both translations are generated using phrase-based translation systems.

The performance of the systems was boosted by using language models built based on different tokens such as word, part-of-speech, and automacally generated word clusters. The difference in word order between German and English is addressed by part-of-speech and syntactic tree-based reordering models. In addition to a discriminative word lexicon, we used hypothesis rescoring using the ListNet algorithm after generating the translation with the phrase-based system. We evaluated the rescoring using only the baseline features as well as using additional computational complex features.

1 Introduction

We describe the KIT systems submitted to the Shared Translation Task of the EMNLP 2015 Tenth Workshop on Statistical Machine Translation. They are phrase-based English→German and German→English systems.

In order to clean a large amount of noisy web-crawled data, we applied a filtering technique using an SVM classifier. Language models are built based on different tokens, such as word, part-of-speech, and automacally generated word clusters. Final systems also include bilingual language models, part-of-speech and syntactic tree-based reordering models as well as a lexicalized reordering model. For language modeling, a data selection strategy is also applied. A discriminative word lexicon using source context information is used for both translation directions. In this evaluation campaign we also show that rescoring using the ListNet algorithm improves the translation performance for both directions.

This paper is organized as follows. In Section 2, we describe the data we used for training the systems. A detailed description of the systems is given in Section 3. Section 4 shows experimental setups and results along with an analysis. Finally, Section 5 concludes this paper.

2 Data

For training data, we use the European Parliament (EPPS), News Commentary (NC) and Common Crawl parallel corpora for both translation directions. For training the language models, we utilize the monolingual target side of the parallel corpora. The News Shuffle data is also used for language modeling. For German→English, we use the Gigaword corpus in addition.

The systems are optimized on the newstest2013 set and tested on the newstest2014 set.

3 System Description

A preprocessing step is applied to the raw data before the actual training. It includes removing excessively long sentences. Sentences with a length mismatch are also filtered out based on a threshold, and special symbols, dates and numbers are normalized. The preprocessing includes smart-casing of the first letter of every sentence. For German→English translation, we apply compound splitting (Koehn and Knight, 2003) on the source side, in order to handle the out-of-vocabulary (OOV) issue of German compound words.

The web-crawled Common Crawl corpus often contains sentence pairs which are not matching. In order to remove such noisy parts of the corpus, we

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 92–97,
Lisboa, Portugal, 17-18 September 2015.

use an SVM classifier for both translation tasks as described in Mediani et al. (2011).

Language models (LM) are built using the SRILM toolkit (Stolcke, 2002) with modified Kneser-Ney smoothing and scored in the decoding process with KenLM (Heafield, 2011). The in-house phrase-based translation system (Vogel, 2003) is used for generating translations. For optimization, we use minimum error rate training (MERT) (Och, 2003; Venugopal et al., 2005). For German→English, the GIZA++ Toolkit (Och and Ney, 2003) is used to generate the word alignment of the parallel corpora. Discriminative word alignment (DWA), as described in Niehues and Vogel (2008), is used for the English→German direction.

We build the phrase tables (PT) using the Moses toolkit (Koehn et al., 2007).

3.1 Word Reordering Models

Reordering rules encode how the words in the source sentence are to be ordered according to the target word order. They are learned automatically based on part-of-speech (POS) as well as syntactic parse tree constituents. In order to learn the rules, we use POS tags (Schmid, 1994) of the source side and the word alignment information. The rules cover short range reorderings (Rottmann and Vogel, 2007) as well as long range reorderings (Niehues and Kolss, 2009).

The differences in word order between German and English can be better addressed by using a tree-based reordering model as shown in Herrmann et al. (2013). The tree-based reordering rules are learned from a word alignment and syntactic parse trees (Rafferty and Manning, 2008; Klein and Manning, 2003) from the source side of the training corpus. The rules encode the information on how to reorder constituents in the syntactic tree of the source sentence.

Before translation, the POS-based and tree-based reordering rules are applied to the each sentence. The variants of differently reordered sentences, including the original order of the sentence, are encoded in a word lattice. The word lattice is then used as an input to the decoder.

Lattice phrase extraction (LPE) (Niehues et al., 2010) is applied on the training corpus, in order to get phrase pairs that match the reordered sentences. In this scheme, we use the reordered sentences to extract the phrases from, instead of the original sentences.

The lexicalized reordering (Koehn et al., 2005) encodes reordering probabilities for each phrase pair. By using the lexicalized reordering model, the reordering orientation of each phrase pair at the phrase boundaries can be determined during decoding. 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 Language Models

In addition to word-based language models, we use different types of non-word language models for each of the systems.

The bilingual language model (Niehues et al., 2011) is designed to increase the bilingual context between source and target words beyond phrase boundaries. Target words and all their aligned source words form bilingual tokens on which a LM is trained. The tokens are then ordered according to the target language word order.

For the English→German system, we use language models based on fine-grained POS tags (Schmid and Laws, 2008). In addition, we use language models based on word classes learned by clustering the words of the corpus using the MKCLS algorithm (Och, 1999). Using such language models, we can generalize better and therefore alleviate the sparsity problem for surface words. In order to build these language models, we replace each word token of the target language corpus by its corresponding POS tag or cluster ID. The n-gram language models are then built on this new corpus consisting of either POS tags or cluster IDs. During decoding, these language models are used as additional models in the log-linear combination.

For the German→English system, the data selection language model is trained on data automatically selected using cross-entropy differences between development sets from previous WMT workshops and the English side of all data, including the filtered crawled data (Moore and Lewis, 2010). We selected the top 10M sentences to train this language model. For building all non-word language models used in this work smoothing is applied.

3.3 Discriminative Word Lexicon

First introduced by Mauser et al. (2009), a discriminative word lexicon (DWL) models the probability of a target word appearing in the translation

given the words of the source sentence. For every target word, a maximum entropy model is trained to determine whether this target word should be in the translated sentence or not using one feature per source word.

Two simplifications of this model are used to improve the translation quality while maintaining the time efficiency as shown in Mediani et al. (2011). First, the score for every phrase pair is calculated before translation. Then we restrict the negative training examples to words that occur within matching phrase pairs.

In this evaluation, the DWL is further extended with n-gram source context features proposed by Niehues and Waibel (2013). In this paper, this model will be referred to as source-context DWL. The source sentence is represented as a bag-of-n-grams, instead of a bag-of-words. By doing so it is possible 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.

In addition to this DWL, we integrated a DWL in the reverse direction in rescoring. We will refer to this model as source DWL. This model predicts the target word for a given source word as described in detail in (Herrmann, 2015).

In a first step, we identify the 20 most frequent translations of each word. Then we build a multi-class classifier to predict the correct translation. For the classifier, we used a binary maximum-entropy classifier[1] trained using the one-against-all approach.

As features for the classifier, we used the previous and following three words. Each word is represented by a continuous vector of 100 dimensions as described in (Mikolov et al., 2013).

Using the predictions, we calculated four additional features. The first two features are the absolute and relative number of words, where the translation predicted by the classifier and the translation in the hypothesis is the same. The third feature is the sum of the word to word translation probabilities predicted by the classifier that occur in the hypothesis. Given the translation used in the hypothesis, we determine their rank in the ranking by the classifier and use the sum of these ranks as the last feature.

[1] http://hal3.name/megam/

3.4 ListNet-based Rescoring

In order to facilitate more complex models like neural network translation models, we rescored the n-best lists. In our experiments we generated 300 best lists for the development and test data respectively. We used the same data to train the rescoring that we have used for optimizing the translation system.

We trained the weights for the log-linear combination used during rescoring using the ListNet algorithm (Cao et al., 2007; Niehues et al., 2015). This technique defines a probability distribution on the permutations of the list based on the scores of the log-linear model and one based on a reference metric. In our experiments we used the BLEU+1 score introduced by Liang et al. (2006). Then we use the cross entropy between both distributions as the loss function for our training.

Using this loss function, we can compute the gradient and use stochastic gradient descent. We used batch updates with ten samples and tuned the learning rate on the development data.

The range of the scores of the different models may greatly differ and many of these values are negative numbers with high absolute value since they are computed as the logarithm of relatively small probabilities. Therefore, we rescale all scores observed on the development data to the range of $[-1, 1]$ prior to rescoring.

3.5 RBM Translation Model

In rescoring, we used an restricted Boltzmann machine (RBM)-based translation model inspired by the work of Devlin et al. (2014).

The model is based on the RBM-based language model introduced in Niehues and Waibel (2012). The RBM models the joint probability of eight target words and a set of attached source words. The set of attached source words is calculated as follows: We first use the source word aligned to the last target word in the 8-gram. If this does not exist, we take the source word aligned to the nearest target word. The set of source words consists then of this source word, its previous five source words and its following five source words.

We create this set of 8 target and 11 source words for every target 8-gram in the parallel corpus and train the model using unigram sampling as described in Niehues et al. (2014). In rescoring, we then calculate the free energy of the RBM given the 8-gram and its source set as input. The

sum of all free energies in the sentence is used as an additional feature for rescoring.

4 Results

In this section, we present a summary of our experiments in the evaluation campaign. Individual components that lead to improvements in the translation performance are described step by step.

The scores are reported in case-sensitive BLEU (Papineni et al., 2002).

4.1 English-German

Table 1 shows the results of our system for English→German translation task.

The baseline system consists of a phrase table derived from DWA, the word-based language models built from different parts of the corpus and POS-based long-range reordering rules. Reordering rules, however, are extracted from the POS-tagged EPPS and NC only, and encoded as word lattices.

The parallel data used to build the word alignments and the PT are EPPS, NC and the filtered Crawl data. Similarly, the data used to train the language models includes the monolingual versions of EPPS, NC and the filtered Crawl data. The BLEU scores of the baseline system over the development and test sets are 19.70 and 19.38, respectively.

The system gains 0.2 points on the development set and 0.13 on the test set in BLEU when adding non-word language models, such as a 4-gram bilingual language model, which is based on bilingual word tokens, two 5-gram POS-based language models and a 4-gram cluster language model. The bilingual language model is trained on the Crawl corpus and the other models are trained on the monolingual parts of all corpora. In case of the cluster language model, MKCLS is used to group of words into 1,000 clusters as mentioned in Section 3.2.

A further improvement can be observed when we apply tree-based and lexicalized reorderings. The improvement is considerable on the development set, gaining 0.6 BLEU points, but the system performs similar on the test set.

Adding source-context DWL helps to improve the score, especially on the test set, with the difference of 0.67 BLEU points compared to the above-mentioned system.

Finally, we use the new ListNet-based rescoring described in Section 3.4 for the log-linear combination of features. By doing so, we improve the translation performance by another 0.8 BLEU points on the test set. This system was submitted to WMT 2015 and used for the translation of the official test set.

System	Dev	Test
Baseline	19.70	19.38
+ Non-word LMs	19.90	19.51
+ Tree + Lex. Reorderings	20.50	19.52
+ Source–context DWL	20.58	20.19
+ ListNet rescoring	19.95	**20.98**

Table 1: Experiments for English→German

4.2 German-English

Table 2 shows the development steps of the German→English translation system.

The baseline system uses EPPS, NC, and filtered web-crawled data for training the translation model. The phrase table is built using GIZA++ word alignment and lattice phrase extraction.

Altogether four language models are used in the baseline system. As described in Section 3.2, we build a cluster language model using the MKCLS algorithm. Words from EPPS, NC, and the filtered crawl data are clustered into $1,000$ different classes. It also includes a language model trained on 10M of selected data from the monolingual corpora. All language models are 4-gram.

The word lattices are generated using short and long-range reordering rules, as well as tree-based reordering rules. A lexicalized reordering model is also included in the baseline system.

The baseline system uses a DWL with source context.

Using the ListNet-based rescoring increased the score on the test set by 0.1 BLEU point. Translation predictions based on source DWL improve the system performance by 0.3 BLEU points. Finally, adding an RBM-based translation model gave another small improvement. This system was used to generate the translation submitted to the evaluation.

5 Conclusion

In this paper, we have described the systems developed for our participation in the Shared Translation Task of the EMNLP 2015 evaluation for

System	Dev	Test
Baseline	28.38	27.77
+ ListNet rescoring	28.00	27.87
+ Source DWL	27.89	28.18
+ RBMTM	27.94	**28.28**

Table 2: Experiments for German→English

English→German and German→English translation. Both translations were generated using a phrase-based translation system which was extended by additional models such as bilingual and cluster-based language models. Discriminative word lexica with source context proved beneficial.

For English→German translation, adding source-context information to guide word choice and using a new method to rescore the translation candidates brought the most improvements.

Rescoring based on ListNet and using source DWL as well as applying an RBM-based translation model helped improve the system performance for German→English translation.

Acknowledgments

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 645452.

References

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136, New York, NY, USA. Acm.

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 52st Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, volume 1, pages 1370–1380, Baltimore, Maryland, USA.

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.

Teresa Herrmann. 2015. Linguistic structure in statistical machine translation.

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.

Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. 2006. An end-to-end discriminative approach to machine translation. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006)*, pages 761–768, Sydney, Australia.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffery Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Workshop at ICLR*, Scottsdale, AZ, USA.

Robert C. Moore and William 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. 2012. Continuous Space Language Models using Restricted Boltzmann Machines. In *Proceedings of the Ninth International Workshop on Spoken Language Translation (IWSLT 2012)*, Hong Kong, HK.

Jan Niehues and Alex 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, Mohammed Mediani, and Alex Waibel. 2010. The Karlsruhe Institute of Technology Translation System for the ACL-WMT 2010. In *Proceedings of the Fifth Workshop on Statistical Machine Translation (WMT 2010)*, Uppsala, Sweden.

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.

Jan Niehues, Alexander Allauzen, François Yvon, and Alex Waibel. 2014. Combining Techniques from Different NN-based Language Models for Machine Translation. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, Vancouver, BC, Canada.

Jan Niehues, Quoc Khanh Do, Alexandre Allauzen, and Alex Waibel. 2015. ListNet-based MT Rescoring. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT 2015)*, Lisboa, Portugal.

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.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 2003)*.

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.