The Karlsruhe Institute of Technology Translation Systems for the WMT 2013

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

Abstract

This paper describes the phrase-based SMT systems developed for our participation in the WMT13 Shared Translation Task. Translations for English↔German and English↔French were generated using a phrase-based translation system which is extended by additional models such as bilingual, fine-grained part-of-speech (POS) and automatic cluster language models and discriminative word lexica (DWL). In addition, we combined reordering models on different sentence abstraction levels.

1 Introduction

In this paper, we describe our systems for the ACL 2013 Eighth Workshop on Statistical Machine Translation. We participated in the Shared Translation Task and submitted translations for English↔German and English↔French using a phrase-based decoder with lattice input.

The paper is organized as follows: the next section 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 System Description

The phrase table is based on a GIZA++ word alignment for the French↔English systems. For the German↔English systems we use a Discriminative Word Alignment (DWA) as described in Niehues and Vogel (2008). For every source phrase only the top 10 translation options are considered during decoding. The SRILM Toolkit (Stolcke, 2002) is used for training SRI language models using Kneser-Ney smoothing.

For the word reordering between languages, we used POS-based reordering models as described in Section 4. In addition to it, tree-based reordering model and lexicalized reordering were added for German↔English systems.

An in-house phrase-based decoder (Vogel, 2003) is used to perform translation. The translation was optimized using Minimum Error Rate Training (MERT) as described in Venugopal et al. (2005) towards better BLEU (Papineni et al., 2002) scores.

2.1 Data

The Europarl corpus (EPPS) and News Commentary (NC) corpus were used for training our translation models. We trained language models for each language on the monolingual part of the training corpora as well as the News Shuffle and the Gigaword corpora. The additional data such as web-crawled corpus, UN and Giga corpora were used after filtering. The filtering work for this data is discussed in Section 3.

For the German↔English systems we use the news-test2010 set for tuning, while the news-test2011 set is used for the French↔English systems. For testing, news-test2012 set was used for all systems.

2.2 Preprocessing

The training data is preprocessed prior to training the system. This includes normalizing special symbols, smart-casing the first word of each sentence and removing long sentences and sentence pairs with length mismatch.

Compound splitting is applied to the German part of the corpus of the German→English system as described in Koehn and Knight (2003).

3 Filtering of Noisy Pairs

The filtering was applied on the corpora which are found to be noisy. Namely, the Giga English-French parallel corpus and the all the new web-crawled data . The operation was performed using

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 104–108,
Sofia, Bulgaria, August 8-9, 2013

an SVM classifier as in our past systems (Mediani et al., 2011). For each pair, the required lexica were extracted from Giza alignment of the corresponding EPPS and NC corpora. Furthermore, for the web-crawled data, higher precision classifiers were trained by providing a larger number of negative examples to the classifier.

After filtering, we could still find English sentences in the other part of the corpus. Therefore, we performed a language identification (LID)-based filtering afterwards (performed only on the French-English corpora, in this participation).

4 Word Reordering

Word reordering was modeled based on POS sequences. For the German↔English system, reordering rules learned from syntactic parse trees were used in addition.

4.1 POS-based Reordering Model

In order to train the POS-based reordering model, probabilistic rules were learned based on the POS tags from the TreeTagger (Schmid and Laws, 2008) of the training corpus and the alignment. As described in Rottmann and Vogel (2007), continuous reordering rules are extracted. This modeling of short-range reorderings was extended so that it can cover also long-range reorderings with non-continuous rules (Niehues and Kolss, 2009), for German↔English systems.

4.2 Tree-based Reordering Model

In addition to the POS-based reordering, we apply a tree-based reordering model for the German↔English translation to better address the differences in word order between German and English. We use the Stanford Parser (Rafferty and Manning, 2008) to generate syntactic parse trees for the source side of the training corpus. Then we use the word alignment between source and target language to learn rules on how to reorder the constituents in a German source sentence to make it match the English target sentence word order better (Herrmann et al., 2013). The POS-based and tree-based reordering rules are applied to each input sentence. The resulting reordered sentence variants as well as the original sentence order are encoded in a word lattice. The lattice is then used as input to the decoder.

4.3 Lexicalized Reordering

The lexicalized reordering model stores the reordering probabilities for each phrase pair. Possible reordering orientations at the incoming and outgoing phrase boundaries are monotone, swap or discontinuous. With the POS- and tree-based reordering word lattices encode different reordering variants. In order to apply the lexicalized reordering model, we store the original position of each word in the lattice. At each phrase boundary at the end, the reordering orientation with respect to the original position of the words is checked. The probability for the respective orientation is included as an additional score.

5 Translation Models

In addition to the models used in the baseline system described above, we conducted experiments including additional models that enhance translation quality by introducing alternative or additional information into the translation modeling process.

5.1 Bilingual Language Model

During the decoding the source sentence is segmented so that the best combination of phrases which maximizes the scores is available. However, this causes some loss of context information at the phrase boundaries. In order to make bilingual context available, we use a bilingual language model (Niehues et al., 2011). In the bilingual language model, each token consists of a target word and all source words it is aligned to.

5.2 Discriminative Word Lexicon

Mauser et al. (2009) introduced the Discriminative Word Lexicon (DWL) into phrase-based machine translation. In this approach, a maximum entropy model is used to determine the probability of using a target word in the translation.

In this evaluation, we used two extensions to this work as shown in (Niehues and Waibel, 2013). First, we added additional features to model the order of the source words better. Instead of representing the source sentence as a bag-of-words, we used a bag-of-n-grams. We used n-grams up to the order of three and applied count filtering to the features for higher order n-grams.

Furthermore, we created the training examples differently in order to focus on addressing errors of the other models of the phrase-based translation

system. We first translated the whole corpus with a baseline system. Then we only used the words that occur in the N-Best List and not in the reference as negative examples instead of using all words that do not occur in the reference.

5.3 Quasi-Morphological Operations

Because of the inflected characteristic of the German language, we try to learn quasi-morphological operations that change the lexical entry of a known word form to the out-of-vocabulary (OOV) word form as described in Niehues and Waibel (2012).

5.4 Phrase Table Adaptation

For the French↔English systems, we built two phrase tables; one trained with all data and the other trained only with the EPPS and NC corpora. This is due to the fact that Giga corpus is big but noisy and EPPS and NC corpus are more reliable. The two models are combined log-linearly to achieve the adaptation towards the cleaner corpora as described in Niehues et al. (2010).

6 Language Models

The 4-gram language models generated by the SRILM toolkit are used as the main language models for all of our systems. For the English↔French systems, we use a good quality corpus as in-domain data to train in-domain language models. Additionally, we apply the POS and cluster language models in different systems. For the German→English system, we build separate language models using each corpus and combine them linearly before the decoding by minimizing the perplexity. Language models are integrated into the translation system by a log-linear combination and receive optimal weights during tuning by the MERT.

6.1 POS Language Models

For the English→German system, we use the POS language model, which is trained on the POS sequence of the target language. The POS tags are generated using the RFTagger (Schmid and Laws, 2008) for German. The RFTagger generates fine-grained tags which include person, gender, and case information. The language model is trained with up to 9-gram information, using the German side of the parallel EPPS and NC corpus, as well as the News Shuffle corpus.

6.2 Cluster Language Models

In order to use larger context information, we use a cluster language model for all our systems. The cluster language model is based on the idea shown in Och (1999). Using the MKCLS algorithm, we cluster the words in the corpus, given a number of classes. Then words in the corpus are replaced with their cluster IDs. Using these cluster IDs, we train n-gram language models as well as a phrase table with this additional factor of cluster ID. Our submitted systems have diversed range of the number of clusters as well as n-gram.

7 Results

Using the models described above we performed several experiments leading finally to the systems used for generating the translations submitted to the workshop. The results are reported as case-sensitive BLEU scores on one reference translation.

7.1 German→English

The experiments for the German to English translation system are summarized in Table 1. The baseline system uses POS-based reordering, DWA with lattice phrase extraction and language models trained on the News Shuffle corpus and Giga corpus separately. Then we added a 5-gram cluster LM trained with 1,000 word classes. By adding a language model using the filtered crawled data we gained 0.3 BLEU on the test set. For this we combined all language models linearly. The filtered crawled data was also used to generate a phrase table, which brought another improvement of 0.85 BLEU. Applying tree-based reordering improved the BLEU score, and the performance had more gain by adding the extended DWL, namely using both bag-of-ngrams and n-best lists. While lexicalized reordering gave us a slight gain, we added morphological operation and gained more improvements.

7.2 English→German

The English to German baseline system uses POS-based reordering and language models using parallel data (EPPS and NC) as shown in Table 2. Gradual gains were achieved by changing alignment from GIZA++ to DWA, adding a bilingual language model as well as a language model based on the POS tokens. A 9-gram cluster-based language model with 100 word classes gave us a

System	Dev	Test
Baseline	24.15	22.79
+ Cluster LM	24.18	22.84
+ Crawled Data LM (Comb.)	24.53	23.14
+ Crawled Data PT	25.38	23.99
+ Tree Rules	25.80	24.16
+ Extended DWL	25.59	24.54
+ Lexicalized Reordering	**26.04**	24.55
+ Morphological Operation	-	**24.62**

Table 1: Translation results for German→English

small gain. Improving the reordering using lexicalized reordering gave us gain on the optimization set. Using DWL let us have more improvements on our test set. By using the filtered crawled data, we gained a big improvement of 0.46 BLEU on the test set. Then we extended the DWL with bag of n-grams and n-best lists to achieve additional improvements. Finally, the best system includes lattices generated using tree rules.

System	Dev	Test
Baseline	17.00	16.24
+ DWA	17.27	16.53
+ Bilingual LM	17.27	16.59
+ POS LM	17.46	16.66
+ Cluster LM	17.49	16.68
+ Lexicalized Reordering	17.57	16.68
+ DWL	17.58	16.77
+ Crawled Data	18.43	17.23
+ Extended DWL	**18.66**	17.57
+ Tree Rules	18.63	**17.70**

Table 2: Translation results for English→German

7.3 French→English

Table 3 reports some remarkable improvements as we combined several techniques on the French→English direction. The baseline system was trained on parallel corpora such as EPPS, NC and Giga, while the language model was trained on the English part of those corpora plus News Shuffle. The newly presented web-crawled data helps to achieve almost 0.6 BLEU points more on test set. Adding bilingual language model and cluster language model does not show a significant impact. Further gains were achieved by the adaptation of in-domain data into general-theme phrase table, bringing 0.15 BLEU better on the test set. When we added the DWL feature, it notably improves the system by 0.25 BLEU points, resulting in our best system.

System	Dev	Test
Baseline	30.33	29.35
+ Crawled Data	30.59	29.93
+ Bilingual and Cluster LMs	30.67	30.01
+ In-Domain PT Adaptation	**31.17**	30.16
+ DWL	31.07	**30.40**

Table 3: Translation results for French→English

7.4 English→French

In the baseline system, EPPS, NC, Giga and News Shuffle corpora are used for language modeling. The big phrase tables tailored EPPC, NC and Giga data. The system also uses short-range reordering trained on EPPS and NC. Adding parallel and filtered crawl data improves the system. It was further enhanced by the integration of a 4-gram bilingual language model. Moreover, the best configuration of 9-gram language model trained on 500 clusters of French texts gains 0.25 BLEU points improvement. We also conducted phrase-table adaptation from the general one into the domain covered by EPPS and NC data and it helps as well. The initial try-out with lexicalized reordering feature showed an improvement of 0.23 points on the development set, but a surprising reduction on the test set, thus we decided to take the system after adaptation as our best English→French system.

System	Dev	Test
Baseline	30.50	27.77
+ Crawled Data	31.05	27.87
+ Bilingual LM	31.23	28.50
+ Cluster LM	31.58	28.75
+ In-Domain PT Adaptation	31.88	**29.12**
+ Lexicalized Reordering	**32.11**	28.98

Table 4: Translation results for English→French

8 Conclusions

We have presented the systems for our participation in the WMT 2013 Evaluation for English↔German and English↔French. All systems use a class-based language model as well as a bilingual language model. Using a DWL with source context improved the translation quality of English↔German systems. Also for these systems, we could improve even more with a tree-based reordering model. Special handling

of OOV words improved German→English system, while for the inverse direction the language model with fine-grained POS tags was helpful. For English↔French, phrase table adaptation helps to avoid using wrong parts of the noisy Giga corpus.

9 Acknowledgements

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

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, June. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *EACL*, Budapest, Hungary.

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: Volume 1 - Volume 1*, EMNLP '09, 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)*.

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

Jan Niehues and Stephan Vogel. 2008. Discriminative Word Alignment via Alignment Matrix Modeling. In *Proc. of Third ACL Workshop on Statistical Machine Translation*, Columbus, USA.

Jan Niehues and Alex Waibel. 2012. Detailed Analysis of different Strategies for Phrase Table Adaptation in SMT. In *Proceedings of the American Machine Translation Association (AMTA)*, San Diego, California, October.

Jan Niehues and Alex Waibel. 2013. An MT Error-driven Discriminative Word Lexicon using Sentence Structure Features. In *Eighth Workshop on Statistical Machine Translation (WMT 2013)*, Sofia, Bulgaria.

Jan Niehues, Mohammed Mediani, Teresa Herrmann, Michael Heck, Christian Herff, and Alex Waibel. 2010. The KIT Translation system for IWSLT 2010. In Marcello Federico, Ian Lane, Michael Paul, and François Yvon, editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 93–98.

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, UK.

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 71–76, Stroudsburg, PA, USA. Association for Computational Linguistics.

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, Ohio.

Kay Rottmann and Stephan Vogel. 2007. Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model. In *TMI*, Skövde, Sweden.

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

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. of ICSLP*, Denver, Colorado, USA.

Ashish Venugopal, Andreas Zollman, and Alex Waibel. 2005. Training and Evaluation Error Minimization Rules for Statistical Machine Translation. In *Workshop on Data-drive Machine Translation and Beyond (WPT-05)*, Ann Arbor, MI.

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