The Karlsruhe Institute of Technology Translation Systems for the WMT 2012

Jan Niehues, Yuqi Zhang, Mohammed Mediani, Teresa Herrmann, Eunah Cho 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 WMT12 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. In addition, we explicitly handle out-of-vocabulary (OOV) words in German, if we have translations for other morphological forms of the same stem. Furthermore, we extended the POS-based reordering approach to also use information from syntactic trees.

1 Introduction

In this paper, we describe our systems for the NAACL 2012 Seventh Workshop on Statistical Machine Translation. We participated in the Shared Translation Task and submitted translations for English↔German and English↔French. We use a phrase-based decoder that can use lattices as input and developed several models that extend the standard log-linear model combination of phrase-based MT. In addition to the POS-based reordering model used in past years, for German-English we extended it to also use rules learned using syntax trees.

The translation model was extended by the bilingual language model and a discriminative word lexicon using a maximum entropy classifier. For the French-English and English-French translation systems, we also used phrase table adaptation to avoid overestimation of the probabilities of the huge, but noisy Giga corpus. In the German-English system, we tried to learn translations for OOV words by exploring different morphological forms of the OOVs with the same lemma.

Furthermore, we combined different language models in the log-linear model. We used word-based language models trained on different parts of the training corpus as well as POS-based language models using fine-grained POS information and language models trained on automatic word clusters.

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

For the French↔English systems the phrase table is based on a GIZA++ word alignment, while the systems for German↔English use a discriminative word alignment as described in Niehues and Vogel (2008). The language models are 4-gram SRI language models using Kneser-Ney smoothing trained by the SRILM Toolkit (Stolcke, 2002).

The problem of word reordering is addressed with POS-based and tree-based reordering models as described in Section 2.3. The POS tags used in the reordering model are obtained using the TreeTagger (Schmid, 1994). The syntactic parse trees are generated using the Stanford Parser (Rafferty and Manning, 2008).

An in-house phrase-based decoder (Vogel, 2003) is used to perform translation. Optimization with

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 349–355,
Montréal, Canada, June 7-8, 2012.

regard to the BLEU score is done using Minimum Error Rate Training as described in Venugopal et al. (2005). During decoding only the top 10 translation options for every source phrase are considered.

2.1 Data

Our translation models were trained on the EPPS and News Commentary (NC) corpora. Furthermore, the additional available data for French and English (i.e. UN and Giga corpora) were exploited in the corresponding systems.

The systems were tuned with the news-test2011 data, while news-test2011 was used for testing in all our systems. We trained language models for each language on the monolingual part of the training corpora as well as the News Shuffle and the Gigaword (version 4) corpora. The discriminative word alignment model was trained on 500 hand-aligned sentences selected from the EPPS corpus.

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 sentences with length mismatch.

For the German parts of the training corpus, in order to obtain a homogenous spelling, we use the hunspell[1] lexicon to map words written according to old German spelling rules to new German spelling rules.

In order to reduce the OOV problem of German compound words, Compound splitting as described in Koehn and Knight (2003) is applied to the German part of the corpus for the German-to-English system.

The Giga corpus received a special preprocessing by removing noisy pairs using an SVM classifier as described in Mediani et al. (2011). The SVM classifier training and test sets consist of randomly selected sentence pairs from the corpora of EPPS, NC, tuning, and test sets. Giving at the end around 16 million sentence pairs.

2.3 Word Reordering

In contrast to modeling the reordering by a distance-based reordering model and/or a lexicalized distortion model, we use a different approach that relies on POS sequences. By abstracting from surface words to POS, we expect to model the reordering more accurately. For German-to-English, we additionally apply reordering rules learned from syntactic parse trees.

2.3.1 POS-based Reordering Model

In order to build the POS-based reordering model, we first learn probabilistic rules from the POS tags of the training corpus and the alignment. Continuous reordering rules are extracted as described in Rottmann and Vogel (2007) to model short-range reorderings. When translating between German and English, we apply a modified reordering model with non-continuous rules to cover also long-range reorderings (Niehues and Kolss, 2009).

2.3.2 Tree-based Reordering Model

Word order is quite different between German and English. And during translation especially verbs or verb particles need to be shifted over a long distance in a sentence. Using discontinuous POS rules already improves the translation tremendously. In addition, we apply a tree-based reordering model for the German-English translation. Syntactic parse trees provide information about the words in a sentence that form constituents and should therefore be treated as inseparable units by the reordering model. For the tree-based reordering model, syntactic parse trees are generated for the whole training corpus. Then the word alignment between the source and target language part of the corpus is used to learn rules on how to reorder the constituents in a German source sentence to make it matches the English target sentence word order better. In order to apply the rules to the source text, POS tags and a parse tree are generated for each sentence. Then the POS-based and tree-based reordering rules are applied. The original order of words as well as the reordered sentence variants generated by the rules are encoded in a word lattice. The lattice is then used as input to the decoder.

For the test sentences, the reordering based on POS and trees allows us to change the word order in the source sentence so that the sentence can be translated more easily. In addition, we build reordering lattices for all training sentences and then extract

[1] http://hunspell.sourceforge.net/

phrase pairs from the monotone source path as well as from the reordered paths.

2.4 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.

2.4.1 Phrase table adaptation

Since the Giga corpus is huge, but noisy, it is advantageous to also use the translation probabilities of the phrase pair extracted only from the more reliable EPPS and News commentary corpus. Therefore, we build two phrase tables for the French↔English system. One trained on all data and the other only trained on the EPPS and News commentary corpus. The two models are then combined using a log-linear combination to achieve the adaptation towards the cleaner corpora as described in (Niehues et al., 2010). The newly created translation model uses the four scores from the general model as well as the two smoothed relative frequencies of both directions from the smaller, but cleaner model. If a phrase pair does not occur in the in-domain part, a default score is used instead of a relative frequency. In our case, we used the lowest probability.

2.4.2 Bilingual Language Model

In phrase-based systems the source sentence is segmented by the decoder according to the best combination of phrases that maximize the translation and language model scores. This segmentation into phrases leads to the loss of context information at the phrase boundaries. Although more target side context is available to the language model, source side context would also be valuable for the decoder when searching for the best translation hypothesis. To make also source language context available we use a bilingual language model, in which each token consists of a target word and all source words it is aligned to. The bilingual tokens enter the translation process as an additional target factor and the bilingual language model is applied to the additional factor like a normal language model. For more details see Niehues et al. (2011).

2.4.3 Discriminative Word Lexica

Mauser et al. (2009) have shown that the use of discriminative word lexica (DWL) can improve the translation quality. For every target word, they trained a maximum entropy model to determine whether this target word should be in the translated sentence or not using one feature per one source word.

When applying DWL in our experiments, we would like to have the same conditions for the training and test case. For this we would need to change the score of the feature only if a new word is added to the hypothesis. If a word is added the second time, we do not want to change the feature value. In order to keep track of this, additional bookkeeping would be required. Also the other models in our translation system will prevent us from using a word too often.

Therefore, we ignore this problem and can calculate the score for every phrase pair before starting with the translation. This leads to the following definition of the model:

$$p(e|f) = \prod_{j=1}^{J} p(e_j|f) \quad (1)$$

In this definition, $p(e_j|f)$ is calculated using a maximum likelihood classifier.

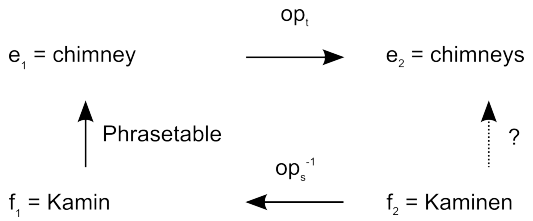
Each classifier is trained independently on the parallel training data. All sentences pairs where the target word e occurs in the target sentence are used as positive examples. We could now use all other sentences as negative examples. But in many of these sentences, we would anyway not generate the target word, since there is no phrase pair that translates any of the source words into the target word.

Therefore, we build a target vocabulary for every training sentence. This vocabulary consists of all target side words of phrase pairs matching a source phrase in the source part of the training sentence. Then we use all sentence pairs where e is in the target vocabulary but not in the target sentences as negative examples. This has shown to have a postive influence on the translation quality (Mediani et al., 2011) and also reduces training time.

2.4.4 Quasi-Morphological Operations for OOV words

Since German is a highly inflected language, there will be always some word forms of a given Ger-

Figure 1: Quasi-morphological operations



man lemma that did not occur in the training data. In order to be able to also translate unseen word forms, we try to learn quasi-morphological operations that change the lexical entry of a known word form to the unknown word form. These have shown to be beneficial in Niehues and Waibel (2011) using Wikipedia[2] titles. The idea is illustrated in Figure 1.

If we look at the data, our system is able to translate a German word *Kamin* (engl. *chimney*), but not the dative plural form *Kaminen*. To address this problem, we try to automatically learn rules how words can be modified. If we look at the example, we would like the system to learn the following rule. If an *"en"* is appended to a German word, as it is done when creating the dative plural form of *Kaminen*, we need to add an *"s"* to the end of the English word in order to perform the same morphological word transformation. We use only rules where the ending of the word has at most 3 letters.

Depending on the POS, number, gender or case of the involved words, the same operation on the source side does not necessarily correspond to the same operation on the target side.

To account for this ambiguity, we rank the different target operation using the following four features and use the best ranked one. Firstly, we should not generate target words that do not exist. Here, we have an advantage that we can use monolingual data to determine whether the word exists. In addition, a target operation that often coincides with a given source operation should be better than one that is rarely used together with the source operation. We therefore look at pairs of entries in the lexicon and count in how many of them the source operation can be applied to the source side and the target operation can be applied to the target side. We then use only operations that occur at least ten times. Furthermore, we use the ending of the source and target word to determine which pair of operations should be used.

[2] http://www.wikipedia.org/

Integration We only use the proposed method for OOVs and do not try to improve translations of words that the baseline system already covers. We look for phrase pairs, for which a source operation op_s exists that changes one of the source words f_1 into the OOV word f_2. Since we need to apply a target operation to one word on the target side of the phrase pair, we only consider phrase pairs where f_1 is aligned to one of the target words of the phrase containing e_1. If a target operation exists given f_1 and op_s, we select the one with the highest rank. Then we generate a new phrase pair by applying op_s to f_1 and op_t to e_1 keeping the original scores from the phrase pairs, since the original and synthesized phrase pair are not directly competing anyway. We do not add several phrase pairs generated by different operations, since we would then need to add the features used for ranking the operations into the MERT. This is problematic, since the operations were only used for very few words and therefore a good estimation of the weights is not possible.

2.5 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 English-French and French-English 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. All language models are integrated into the translation system by a log-linear combination and received optimal weights during tuning by the MERT.

2.5.1 POS Language Models

The POS language model is trained on the POS sequences of the target language. In this evaluation, the POS language model is applied for the English-German system. We expect that having additional information in form of probabilities of POS sequences should help especially in case of the rich morphology of German. The POS tags are generated with the RFTagger (Schmid and Laws, 2008) for German, which produces fine-grained tags that include person, gender and case information. We

use a 9-gram language model on the News Shuffle corpus and the German side of all parallel corpora. More details and discussions about the POS language model can be found in Herrmann et al. (2011).

2.5.2 Cluster Language Models

The cluster language model follows a similar idea as the POS language model. Since there is a data sparsity problem when we substitute words with the word classes, it is possible to make use of larger context information. In the POS language model, POS tags are the word classes. Here, we generated word classes in a different way. First, we cluster the words in the corpus using the MKCLS algorithm (Och, 1999) given a number of classes. Second, we replace the words in the corpus by their cluster IDs. Finally, we train an n-gram language model on this corpus consisting of cluster IDs. Generally, all cluster language models used in our systems are 5-gram.

3 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 following sections describe the experiments for the individual language pairs and show the translation results. The results are reported as case-sensitive BLEU scores (Papineni et al., 2002) on one reference translation.

3.1 German-English

The experiments for the German-English translation system are summarized in Table 1. The Baseline system uses POS-based reordering, discriminative word alignment and a language model trained on the News Shuffle corpus. By adding lattice phrase extraction small improvements of the translation quality could be gained.

Further improvements could be gained by adding a language model trained on the Gigaword corpus and adding a bilingual and cluster-based language model. We used 50 word classes and trained a 5-gram language model. Afterwards, the translation quality was improved by also using a discriminative word lexicon. Finally, the best system was achieved by using Tree-based reordering and using special treatment for the OOVs. This system generates a BLEU score of 22.31 on the test data. For the last two systems, we did not perform new optimization runs.

System	Dev	Test
Baseline	23.64	21.32
+ Lattice Phrase Extraction	23.76	21.36
+ Gigaward Language Model	24.01	21.73
+ Bilingual LM	24.19	21.91
+ Cluster LM	24.16	22.09
+ DWL	24.19	22.19
+ Tree-based Reordering	-	22.26
+ OOV	-	**22.31**

Table 1: Translation results for German-English

3.2 English-German

The English-German baseline system uses also POS-based reordering, discriminative word alignment and a language model based on EPPS, NC and News Shuffle. A small gain could be achieved by the POS-based language model and the bilingual language model. Further gain was achieved by using also a cluster-based language model. For this language model, we use 100 word classes and trained a 5-gram language model. Finally, the best system uses the discriminative word lexicon.

System	Dev	Test
Baseline	17.06	15.57
+ POSLM	17.27	15.63
+ Bilingual LM	17.40	15.78
+ Cluster LM	17.77	16.06
+ DWL	**17.75**	**16.28**

Table 2: Translation results for English-German

3.3 English-French

Table 3 summarizes how our English-French system evolved. The baseline system here was trained on the EPPS, NC, and UN corpora, while the language model was trained on all the French part of the parallel corpora (including the Giga corpus). It also uses short-range reordering trained on EPPS and NC. This system had a BLEU score of around 26.7. The Giga parallel data turned out to be quite

beneficial for this task. It improves the scores by more than 1 BLEU point. More importantly, additional language models boosted the system quality: around 1.8 points. In fact, three language models were log-linearly combined: In addition to the aforementioned, two additional language models were trained on the monolingual sets (one for News and one for Gigaword). We could get an improvement of around 0.2 by retraining the reordering rules on EPPS and NC only, but using Giza alignment from the whole data. Adapting the translation model by using EPPS and NC as in-domain data improves the BLEU score by only 0.1. This small improvement might be due to the fact that the news domain is very broad and that the Giga corpus has already been carefully cleaned and filtered. Furthermore, using a bilingual language model enhances the BLEU score by almost 0.3. Finally, incorporating a cluster language model adds an additional 0.1 to the score. This leads to a system with 30.58.

System	Dev	Test
Baseline	24.96	26.67
+ GigParData	26.12	28.16
+ Big LMs	29.22	29.92
+ All Reo	29.14	30.10
+ PT Adaptation	29.15	30.22
+ Bilingual LM	29.17	30.49
+ Cluster LM	**29.08**	**30.58**

Table 3: Translation results for English-French

3.4 French-English

The development of our system for the French-English direction is summarized in Table 4. The baseline system for this direction was trained on the EPPS, NC, UN and Giga parallel corpora, while the language model was trained on the French part of the parallel training corpora. The baseline system includes the POS-based reordering model with short-range rules. The largest improvement of 1.7 BLEU score was achieved by the integration of the bigger language models which are trained on the English version of News Shuffle and the Gigaword corpus (v4). We did not add the language models from the monolingual English version of EPPS and NC data, since the experiments have shown that they did not provide improvement in our system. The second largest improvement came from the domain adaptation that includes an in-domain language model and adaptations to the phrase extraction. The BLEU score has improved about 1 BLEU in total. The in-domain data we used here are parallel EPPS and NC corpus. Further gains were obtained by augmenting the system with a bilingual language model adding around 0.2 BLEU to the previous score. The submitted system was obtained by adding the cluster 5-gram language model trained on the News Shuffle corpus with 100 clusters and thus giving 30.25 as the final score.

System	Dev	Test
Baseline	25.81	27.15
+ Indomain LM	26.17	27.91
+ PT Adaptation	26.33	28.11
+ Big LMs	28.90	29.82
+ Bilingual LM	29.14	30.09
+ Cluster LM	**29.31**	**30.25**

Table 4: Translation results for French-English

4 Conclusions

We have presented the systems for our participation in the WMT 2012 Evaluation for English↔German and English↔French. In all systems we could improve by using a class-based language model. Furthermore, the translation quality could be improved by using a discriminative word lexicon. Therefore, we trained a maximum entropy classifier for every target word. For English↔French, adapting the phrase table helps to avoid using wrong parts of the noisy Giga corpus. For the German-to-English system, we could improve the translation quality additionally by using a tree-based reordering model and by special handling of OOV words. For the inverse direction we could improve the translation quality by using a 9-gram language model trained on the fine-grained POS tags.

Acknowledgments

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

References

Teresa Herrmann, Mohammed Mediani, Jan Niehues, and Alex Waibel. 2011. The karlsruhe institute of technology translation systems for the wmt 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 379–385, Edinburgh, Scotland, July. 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. 2011. Using wikipedia to translate domain-specific terms in smt. In *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*.

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*.

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.

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

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.