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 include bilingual language models, reordering models based on part-of-speech tags and syntactic parse 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 modeling. 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 participated 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 trained on hand-aligned data as described in Niehues and Vogel (2008). The phrase table (PT) is built using the Moses toolkit (Koehn et al., 2007). The phrase 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.

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>15.63</td>
<td>27.61</td>
</tr>
<tr>
<td>+ Big LMs</td>
<td>16.56</td>
<td>29.02</td>
</tr>
<tr>
<td>+ PT Adaptation</td>
<td>16.77</td>
<td>29.32</td>
</tr>
<tr>
<td>+ Bilingual + Cluster LM</td>
<td>16.87</td>
<td>29.64</td>
</tr>
<tr>
<td>+ Lexicalized Reordering</td>
<td>16.92</td>
<td>30.17</td>
</tr>
<tr>
<td>+ Source DWL</td>
<td>17.28</td>
<td><strong>30.19</strong></td>
</tr>
</tbody>
</table>

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 preproces-
ing 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.

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>30.16</td>
<td>30.70</td>
</tr>
<tr>
<td>+ LM Adaptation</td>
<td>30.58</td>
<td>30.94</td>
</tr>
<tr>
<td>+ PT Adaptation</td>
<td>30.69</td>
<td>31.14</td>
</tr>
<tr>
<td>+ Bilingual + Cluster LM</td>
<td>30.85</td>
<td>31.38</td>
</tr>
<tr>
<td>+ Lexicalized Reordering</td>
<td>31.14</td>
<td>31.46</td>
</tr>
<tr>
<td>+ Source DWL</td>
<td>31.19</td>
<td>31.54</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>16.64</td>
<td>18.60</td>
</tr>
<tr>
<td>+ Cluster LM</td>
<td>16.76</td>
<td>18.66</td>
</tr>
<tr>
<td>+ Common Crawl Data</td>
<td>17.27</td>
<td>19.66</td>
</tr>
<tr>
<td>+ LPE + Lexicalized Reordering</td>
<td>17.45</td>
<td>19.75</td>
</tr>
<tr>
<td>+ Tree Rules</td>
<td>17.53</td>
<td>19.85</td>
</tr>
<tr>
<td>+ Shorter n-grams</td>
<td>17.55</td>
<td>19.92</td>
</tr>
<tr>
<td>+ Source DWL + Big BiLM</td>
<td>17.82</td>
<td>20.21</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>24.40</td>
<td>26.34</td>
</tr>
<tr>
<td>+ Tree Rules</td>
<td>24.71</td>
<td>26.86</td>
</tr>
<tr>
<td>+ Lexicalized Reordering</td>
<td>24.89</td>
<td>26.93</td>
</tr>
<tr>
<td>+ LM Data Selection</td>
<td>24.96</td>
<td>27.03</td>
</tr>
<tr>
<td>+ Source DWL</td>
<td>25.32</td>
<td>27.53</td>
</tr>
<tr>
<td>+ Morphological Operations</td>
<td>-</td>
<td>27.53</td>
</tr>
</tbody>
</table>

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


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 Statistical Translation, Atlanta, Georgia, USA.


