The Universität Karlsruhe Translation System for the EACL-WMT 2009

Jan Niehues, Teresa Herrmann, Muntsin Kolss and Alex Waibel
Universität Karlsruhe (TH)
Karlsruhe, Germany
{jniehues,therrman,kolss,waibel}@ira.uka.de

Abstract

In this paper we describe the statistical machine translation system of the Universität Karlsruhe developed for the translation task of the Fourth Workshop on Statistical Machine Translation. The state-of-the-art phrase-based SMT system is augmented with alternative word reordering and alignment mechanisms as well as optional phrase table modifications. We participate in the constrained condition of German-English and English-German as well as in the constrained condition of French-English and English-French.

1 Introduction

This paper describes the statistical MT system used for our participation in the WMT'09 Shared Translation Task and the particular language-pair-dependent variations of the system. We use standard alignment and training tools and a phrase-based SMT decoder for creating state-of-the-art MT systems for our contribution in the translation directions English-German, German-English, English-French and French-English.

Depending on the language pair, the baseline system is augmented with part-of-speech (POS)-based short-range and long-range word reordering models, discriminative word alignment (DWA) and several modifications of the phrase table. Experiments with different system variants were conducted including some of those additional system components. Significantly better translation results could be achieved compared to the baseline results.

An overview of the system will follow in Section 2, which describes the baseline architecture, followed by descriptions of the additional system components. Translation results for the different languages and system variants are presented in Section 5.

2 Baseline System

The core of our system is the STTK decoder (Vogel, 2003), a phrase-based SMT decoder with a local reordering window of 2 words. The decoder generates a translation for the input text or word lattice by searching translation model and language model for the hypothesis that maximizes phrase translation probabilities and target language probabilities. The translation model, i.e. the SMT phrase table is created during the training phase by a modified version of the Moses Toolkit (Koehn et al., 2007) applying GIZA++ for word alignment. Language models are built using the SRILM Toolkit. The POS-tags for the reordering models were generated with the TreeTagger (Schmid, 1994) for all languages.

2.1 Training, Development and Test Data

We submitted translations for the English-German, German-English, English-French and French-English tasks. All systems were trained on the Europarl and News Commentary corpora using the Moses Toolkit and apply 4-gram language models created from the respective monolingual News corpora. All feature weights are automatically determined and optimized with respect to BLEU via MERT (Venugopal et al., 2005). For development and testing we used data provided by the WMT'09, news-dev2009a and news-dev2009b, consisting of 1026 sentences each.

3 Word Reordering Model

One part of our system that differs from the baseline system is the reordering model. To account for the different word orders in the languages, we used the POS-based reordering model presented in Rottmann and Vogel (2007). This model learns rules from a parallel text to reorder the source side. The aim is to generate a reordered source side that can be translated in a more monotone way.

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 80–84,
Athens, Greece, 30 March – 31 March 2009.

In this framework, first, reordering rules are extracted from an aligned parallel corpus and POS information is added to the source side. These rules are of the form *VVIMP VMFIN PPER → PPER VMFIN VVIMP* and describe how the source side has to be reordered to match the target side. Then the rules are scored according to their relative frequencies.

In a preprocessing step to the actual decoding different reorderings of the source sentences are encoded in a word lattice. Therefore, for all reordering rules that can be applied to a sentence the resulting reorderings are added to the lattice if the score is better than a given threshold. The decoding is then performed on the resulting word lattice.

This approach does model the reordering well if only short-range reorderings occur. But especially when translating from and to German, there are also long-range reorderings that require the verb to be shifted nearly across the whole sentence. During this shift of the verb, the rest of the sentence remains mainly unchanged. It does not matter which words are in between, since they are moved as a whole. Furthermore, rules including an explicit sequence of POS-tags spanning the whole sentence would be too specific. A lot more rules would be needed to cover long-range reorderings with each rule being applicable only very sparsely. Therefore, we model long-range reordering by generalizing over the unaffected sequences and introduce rules with gaps. (For more details see Niehues and Kolss (2009)). These are learned in a way similar to the other type of reordering rules described above, but contain a gap representing one or several arbitrary words. It is, for example, possible to have the following rule *VAFIN * VVPP → VAFIN VVPP **, which puts both parts of the German verb next to each other.

4 Translation Model

The translation models of all systems we submitted differ in some parts from the baseline system. The main changes done will be described in this section.

4.1 Word Alignment

The baseline method for creating the word alignment is to create the GIZA++ alignments in both directions and then to combine both alignments using a heuristic, e.g. grow-diag-final-and heuristic, as provided by the Moses Toolkit. In some of the submitted systems we used a discriminative word alignment model (*DWA*) to generate the alignments as described in Niehues and Vogel (2008) instead. This model is trained on a small amount of hand-aligned data and uses the lexical probability as well as the fertilities generated by the GIZA++ Toolkit and POS information. We used all local features, the GIZA and indicator fertility features as well as first order features for 6 directions. The model was trained in three steps, first using the maximum likelihood optimization and afterwards it was optimized towards the alignment error rate. For more details see Niehues and Vogel (2008).

4.2 Phrase Table Smoothing

The relative frequencies of the phrase pairs are a very important feature of the translation model, but they often overestimate rare phrase pairs. Therefore, the raw relative frequency estimates found in the phrase translation tables are smoothed by applying modified Kneser-Ney discounting as described in Foster et al. (2006).

4.3 Lattice Phrase Extraction

For the test sentences the POS-based reordering allows us to change the word order in the source sentence, so that the sentence can be translated more easily. But this approach does not reorder the training sentences. This may cause problems for phrase extraction, especially for long-range reorderings. For example, if the English verb is aligned to both parts of the German verb, this phrase can not be extracted, since it is not continuous on the German side. In the case of German as source language, the phrase could be extracted if we also reorder the training corpus.

Therefore, we build lattices that encode the different reorderings for every training sentence. Then we can not only extract phrase pairs from the monotone source path, but also from the reordered paths. So it would be possible to extract the example mentioned before, if both parts of the verb were put together by a reordering rule. To limit the number of extracted phrase pairs, we extract a source phrase only once per sentence even if it may be found on different paths. Furthermore, we do not use the weights in the lattice.

If we use the same rules as for the test sets, the lattice would be so big that the number of extracted phrase pairs would be still too high. As mentioned before, the word reordering is mainly

a problem at the phrase extraction stage if one word is aligned to two words which are far away from each other in the sentence. Therefore, the short-range reordering rules do not help much in this case. So, only the long-range reordering rules were used to generate the lattice for the training corpus. This already leads to an increase of the number of source phrases in the filtered phrase table from 724K to 971K. The number of phrase pairs grows from 5.1M to 6.7M.

4.4 Phrase Table Adaption

For most of the different tasks there was a huge amount of parallel out-of-domain training data available, but only a much smaller amount of in-domain training data. Therefore, we tried to adapt our system to the in-domain data. We want to make use of the big out-of-domain data, but do not want to lose the information encoded in the in-domain data.

To achieve this, we built an additional phrase table trained only on the in-domain data. Since the word alignment does not depend heavily on the domain we used the same word alignment. Then we combined both phrase tables in the following way. A phrase pair with features θ from the first phrase table is added to the combined one with features $< \theta, 1 >$, where 1 is a vector of ones with length equal to the number of features in the other phrase table. The phrase pairs of the other phrase table were added with the features $< 1, \theta >$.

5 Results

We submitted system translations for the English-German, German-English, English-French and French-English task. Their performance is measured applying the BLEU metric. All BLEU scores are computed on the lower-cased translations.

5.1 English-German

The system translating from English to German was trained on the data described in Section 2.1. The first system already uses the POS-based reordering model for short-range reorderings. The results of the different systems are shown in Table 1.

We could improve the translation quality on the test set by using the smoothed relative frequencies in the phrase table as described before and by adapting the phrase table. Then we used the discriminative word alignment to generate a new word alignment. For the training of the model we used 500 hand-aligned sentences from the Europarl corpus. By training a translation model based on this word alignment we could improve the translation quality further. At last we added the model for long-range reorderings, which performs best on the test set.

The improvement achieved by smoothing is significant at a level of 5%, the remaining changes are not significant on their own. In all language pairs, the problem occurs that some features do not lead to an improvement on the development set, but on the test set. One reason for this may be that the development set is quite small.

Table 1: Translation results for English-German (BLEU Score)

System	Dev	Test
Short-range	13.96	14.99
+ Smoothing	14.36	15.38
+ Adaptation	13.96	15.44
+ Discrim. WA	14.45	15.61
+ Long-range reordering	14.58	**15.70**

5.2 German-English

The German-English system was trained on the same data as the English-German except that we perform compound splitting as an additional preprocessing step. The compound splitting was done with the frequency-based method described in Koehn et al. (2003). For this language direction, the initial system already uses phrase table smoothing, adaptation and discriminative word alignment, in addition to the techniques of the English-German baseline system. The results are shown in Table 2.

For this language pair, we could improve the translation quality, first, by adding the long-range reordering model. Further improvements could be achieved by using lattice phrase extraction as described before.

5.3 English-French

For creating the English-French translations, first, the baseline system as described in Section 2 was used. This baseline was then augmented with phrase table smoothing, short-range word reordering and phrase table adaptation as described above. In addition, the adapted phrase table was

Table 2: Translation results for German-English (BLEU Score)

System	Dev	Test
Initial System	20.52	22.01
+ Long-range reordering	21.04	22.36
+ Lattice phrase extraction	20.69	**22.64**

postprocessed such that phrase table entries include the same amount of punctuation marks, especially quotation marks, in both source and target phrase. In contrast to the English↔German language pairs, the word reordering required in English↔French translations are restricted to rather local word shifts which can be covered by the short-range reordering feature. Applying additional long-range reordering is scarcely expected to yield further improvements for these language pairs and was not applied specifically in this task. Table 3 shows the results of the system variants.

Table 3: Translation results for English-French (BLEU Score)

System	Dev	Test
Baseline	20.97	20.87
+ Smoothing	21.42	21.32
+ Short-range reordering	20.79	**22.26**
+ Adaptation	21.05	21.97
+ cleanPT	21.50	21.98

Both on development and test set, smoothing the probabilities in the phrase table resulted in an increase of nearly 0.5 BLEU points. Applying short-range word reordering did not lead to an improvement on the development set. However, the increase in BLEU on the test set is substantial. The opposite is the case when adapting the phrase table: While phrase table adaptation improves the translation quality on the development set, adaptation leads to lower scores on the test set.

Thus, the system configuration that performed best on the test set applies phrase table smoothing and short-range word reordering. For creating the translations for our submission, this configuration was used.

5.4 French-English

For the French-English task, similar experiments have been conducted. With respect to the baseline system, improvements in translation quality could be measured when applying phrase table smoothing. An increase of 0.43 BLEU points was achieved using short-range word reordering. Additional experiments with adapting the phrase table to the domain of the test set led to further improvement. Submissions for the shared task were created using the system including all mentioned features.

Table 4: Translation results for French-English (BLEU Score)

System	Dev	Test
Baseline	21.29	22.41
+ Smoothing	21.55	22.59
+ Short-range reordering	22.55	23.02
+ Adaptation	21.72	23.20
+ cleanPT	22.60	**23.21**

6 Conclusions

We have presented our system for the WMT'09 Shared Translation Task. The submissions for the language pairs English-German, German-English, English-French and French-English have been created by the STTK decoder applying different additional methods for each individual language pair to enhance translation quality.

Word reordering models covering short-range reordering for the English↔French and English↔German and long-range reordering for English↔German respectively proved to result in better translations.

Smoothing the phrase probabilities in the phrase table also increased the scores in all cases, while adapting the phrase table to the test domain only showed a positive influence on translation quality in some of our experiments. Further tuning of the adaptation procedure could help to clarify the benefit of this method.

Using discriminative word alignment as an alternative to performing word alignment with GIZA++ did also improve the systems translating between English and German. Future experiments will be conducted applying discriminative word alignment also in the English↔French systems.

Acknowledgments

This work was partly supported by Quaero Programme, funded by OSEO, French State agency for innovation.

References

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *Proc. of Empirical Methods in Natural Language Processing*. Sydney, Australia.

Philipp Koehn, Franz Josef Och and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *HLT/NAACL 2003*. Edmonton, Canada.

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 *Proc. of Second ACL Workshop on Statistical Machine Translation*. Prague, Czech Republic.

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

Jan Niehues and Muntsin Kolss. 2009. A POS-Based Model for Long-Range Reorderings in SMT. In *Proc. of Forth ACL Workshop on Statistical Machine Translation*. Athens, Greece.

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

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

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

Stephan Vogel. 2003. SMT Decoder Dissected: Word Reordering. In *NLP-KE'03*. Beijing, China.