The Karlsruhe Institute for Technology Translation System for the ACL-WMT 2010

Jan Niehues, Teresa Herrmann, Mohammed Mediani and Alex Waibel

Karlsruhe Insitute of Technolgy

Karlsruhe, Germany

`firstname.lastname@kit.edu`

Abstract

This paper describes our phrase-based Statistical Machine Translation (SMT) system for the WMT10 Translation Task. We submitted translations for the German to English and English to German translation tasks. Compared to state-of-the-art phrase-based systems we preformed additional preprocessing and used a discriminative word alignment approach. The word reordering was modeled using POS information and we extended the translation model with additional features.

1 Introduction

In this paper we describe the systems that we built for our participation in the Shared Translation Task of the ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. Our translations are generated using a state-of-the-art phrase-based translation system and applying different extensions and modifications including Discriminative Word Alignment, a POS-based reordering model and bilingual language models using POS and stem information.

Depending on the source and target languages, the proposed models differ in their benefit for the translation task and also expose different correlative effects. The Sections 2 to 4 introduce the characteristics of the baseline system and the supplementary models. In Section 5 we present the performance of the system variants applying the different models and chose the systems used for creating the submissions for the English-German and German-English translation task. Section 6 draws conclusions and suggests directions for future work.

2 Baseline System

The baseline systems for the translation directions German-English and English-German are both developed using Discriminative Word Alignment (Niehues and Vogel, 2008) and the Moses Toolkit (Koehn et al., 2007) for extracting phrase pairs and generating the phrase table from the discriminative word alignments. The difficult reordering between German and English was modeled using POS-based reordering rules. These rules were learned using a word-aligned parallel corpus. The POS tags for the reordering models are generated using the TreeTagger (Schmid, 1994) for all languages.

Translation is performed by the STTK Decoder (Vogel, 2003) and all systems are optimized towards BLEU using Minimum Error Rate Training as proposed in Venugopal et al. (2005).

2.1 Training, Development and Test Data

We used the data provided for the WMT for training, optimizing and testing our systems: Our training corpus consists of Europarl and News Commentary data, for optimization we use newstest2008 as development set and newstest2009 as test set.

The baseline language models are trained on the target language part of the Europarl and News Commentary corpora. Additional, bigger language models were trained on monolingual corpora. For both systems the News corpus was used while an English language model was also trained on the even bigger Gigaword corpus.

2.2 Preprocessing

The training data was preprocessed before used for training. In this step different normalizations were done like mapping different types of quotes. In the end the first word of every sentence was smartcased.

For the German text, additional preprocessing steps were applied. First, the older German data uses the old German orthography whereas the newer parts of the corpus use the new German orthography. We tried to normalize the text by converting the whole text to the new German orthography. In a first step, we search for words that are only correct according to the old writing rules. Therefore, we selected all words in the corpus, that are correct according to the hunspell lexicon[1] using the old rules, but not correct according to the hunspell lexicon using the new rules. In a second step we tried to find the correct spelling according to the new rules. We first applied rules describing how words changed from one spelling system to the other, for example replacing 'ß' by 'ss'. If the new word is a correct word according to the hunspell lexicon using the new spelling rules, we map the words.

When translating from German to English, we apply compound splitting as described in Koehn and Knight (2003) to the German corpus.

As a last preprocessing step we remove sentences that are too long and empty lines to obtain the final training corpus.

3 Word Reordering Model

Reordering was applied on the source side prior to decoding through the generation of lattices encoding possible reorderings of each source sentence that better match the word sequence in the target language. These possible reorderings were learned based on the POS of the source language words in the training corpus and the information about alignments between source and target language words in the corpus. For short-range reorderings, continuous reordering rules were applied to the test sentences (Rottmann and Vogel, 2007). To model the long-range reorderings between German and English, different types of non-continuous reordering rules were applied depending on the translation direction. (Niehues and Kolss, 2009). When translating from English to German, most of the changes in word order consist in a shift to the right while typical word shifts in German to English translations take place in the reverse direction.

4 Translation Model

The translation model was trained on the parallel corpus and the word alignment was generated by a discriminative word alignment model, which is described below. The phrase table was trained using the Moses training scripts, but for the German to English system we used a different phrase extraction method described in detail in Section 4.2. In addition, we applied phrase table smoothing as described in Foster et al. (2006). Furthermore, we extended the translation model by additional features for unaligned words and introduced bilingual language models.

4.1 Word Alignment

In most phrase-based SMT systems the heuristic grow-diag-final-and is used to combine the alignments generated by GIZA++ from both directions. Then these alignments are used to extract the phrase pairs.

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 PGIZA++[2] 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 maximum likelihood optimization and afterwards it was optimized towards the alignment error rate. For more details see Niehues and Vogel (2008).

4.2 Lattice Phrase Extraction

In translations from German to English, we often have the case that the English verb is aligned to both parts of the German verb. Since this phrase pair is not continuous on the German side, it cannot be extracted. The phrase could be extracted, if we also reorder the training corpus.

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. If we apply this also to the training sentences, we would be able to extract the phrase pairs for originally discontinuous phrases and could apply them during translation of the reordered test sentences.

Therefore, we build lattices that encode the different reorderings for every training sentence, as described in Niehues et al. (2009). 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 used the same rules as for reordering 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 lattices for the training corpus.

4.3 Unaligned Word Feature

Guzman et al. (2009) analyzed the role of the word alignment in the phrase extraction process. To better model the relation between word alignment and the phrase extraction process, they introduced two new features into the log-linear model. One feature counts the number of unaligned words on the source side and the other one does the same for the target side. Using these additional features they showed improvements on the Chinese to English translation task. In order to investigate the impact on closer related languages like English and German, we incorporated those two features into our systems.

4.4 Bilingual Word language model

Motivated by the improvements in translation quality that could be achieved by using the n-gram based approach to statistical machine translation, for example by Allauzen et al. (2009), we tried to integrate a bilingual language model into our phrase-based translation system.

To be able to integrate the approach easily into a standard phrase-based SMT system, a token in the bilingual language model is defined to consist of a target word and all source words it is aligned to. The tokens are ordered according to the target language word order. Then the additional tokens can be introduced into the decoder as an additional target factor. Consequently, no additional implementation work is needed to integrate this feature.

If we have the German sentence *Ich bin nach Hause gegangen* with the English translation *I went home*, the resulting bilingual text would look like this: *I_Ich went_bin_gegangen home_Hause*.

As shown in the example, one problem with this approach is that unaligned source words are ignored in the model. One solution could be to have a second bilingual text ordered according to the source side. But since the target sentence and not the source sentence is generated from left to right during decoding, the integration of a source side language model is more complex. Therefore, as a first approach we only used a language model based on the target word order.

4.5 Bilingual POS language model

The main advantage of POS-based information is that there are less data sparsity problems and therefore a longer context can be considered. Consequently, if we want to use this information in the translation model of a phrase-based SMT system, the POS-based phrase pairs should be longer than the word-based ones. But this is not possible in many decoders or it leads to additional computation overhead.

If we instead use a bilingual POS-based language model, the context length of the language model is independent from the other models. Consequently, a longer context can be considered for the POS-based language model than for the word-based bilingual language model or the phrase pairs.

Instead of using POS-based information, this approach can also be applied with other additional linguistic word-level information like word stems.

5 Results

We submitted translations for English-German and German-English for the Shared Translation Task. In the following we present the experiments we conducted for both translation directions applying the aforementioned models and extensions to the baseline systems. The performance of each individual system configuration was measured applying the BLEU metric. All BLEU scores are calculated on the lower-cased translation hypotheses. The individual systems that were used to create the submission are indicated in bold.

5.1 English-German

The baseline system for English-German applies short-range reordering rules and discriminative word alignment. The language model is trained on the News corpus. By expanding the coverage of the rules to enable long-range reordering, the score on the test set could be slightly improved. We then combined the target language part of the Europarl and News Commentary corpora with the News corpus to build a bigger language model which resulted in an increase of 0.11 BLEU points on the development set and an increase of 0.25 points on the test set. Applying the bilingual language model as described above led to 0.04 points improvement on the test set.

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

System	Dev	Test
Baseline	15.30	15.40
+ Long-range Reordering	15.25	15.44
+ EPNC LM	15.36	15.69
+ bilingual Word LM	**15.37**	**15.73**
+ bilingual POS LM	15.42	15.67
+ unaligned Word Feature	15.65	15.66
+ bilingual Stem LM	15.57	15.74

This system was used to create the submission to the Shared Translation Task of the WMT 2010. After submission we performed additional experiments which only led to inconclusive results. Adding the bilingual POS language model and introducing the unaligned word feature to the phrase table only improved on the development set, while the scores on the test set decreased. A third bilingual language model based on stem information again only showed noteworthy effects on the development set.

5.2 German-English

For the German to English translation system, the baseline system already uses short-range reordering rules and the discriminative word alignment. This system applies only the language model trained on the News corpus. By adding the possibility to model long-range reorderings with POS-based rules, we could improve the system by 0.6 BLEU points. Adding the big language model using also the English Gigaword corpus we could improve by 0.3 BLEU points. We got an additional improvement by 0.1 BLEU points by adding lattice phrase extraction.

Both the word-based and POS-based bilingual language model could improve the translation quality measured in BLEU. Together they improved the system performance by 0.2 BLEU points.

The best results could be achieved by using also the unaligned word feature for source and target words leading to the best performance on the test set (22.09).

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

System	Dev	Test
Baseline	20.94	20.83
+ Long-range Reordering	21.52	21.43
+ Gigaword LM	21.90	21.71
+ Lattice Phrase Extraction	21.94	21.81
+ bilingual Word LM	21.94	21.87
+ bilingual POS LM	22.02	22.05
+ unaligned Word Feature	**22.09**	**22.09**

6 Conclusions

For our participation in the WMT 2010 we built translation systems for German to English and English to German. We addressed to the difficult word reordering when translating from or to German by using POS-based reordering rules during decoding and by using lattice-based phrase extraction during training. By applying those methods we achieved substantially better results for both translation directions.

Furthermore, we tried to improve the translation quality by introducing additional features to the translation model. On the one hand we included bilingual language models based on different word factors into the log-linear model. This led to very slight improvements which differed also with respect to language and data set. We will investigate in the future whether further improvements are achievable with this approach. On the other hand we included the unaligned word feature which has been applied successfully for other language pairs. The improvements we could gain with this method are not as big as the ones reported for other languages, but still the performance of our systems could be improved using this feature.

References

Alexandre Allauzen, Josep Crego, Aurélien Max, and François Yvon. 2009. LIMSI's statistical translation system for WMT'09. In *Fourth Workshop on Statistical Machine Translation (WMT 2009)*, Athens, Greece.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, Sydney, Australia.

Francisco Guzman, Qin Gao, and Stephan Vogel. 2009. Reassessment of the Role of Phrase Extraction in PBSMT. In *MT Summit XII*, Ottawa, Ontario, Canada.

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

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 *ACL 2007, Demonstration Session*, Prague, Czech Republic, June 23.

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, Teresa Herrmann, Muntsin Kolss, and Alex Waibel. 2009. The Universität Karlsruhe Translation System for the EACL-WMT 2009. In *Fourth Workshop on Statistical Machine Translation (WMT 2009)*, 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 *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.