**The Karlsruhe Institute of Technology Translation Systems for the WMT 2011**

**Teresa Herrmann, Mohammed Mediani, Jan Niehues and Alex Waibel**

Karlsruhe Institute of Technology Karlsruhe, Germany [firstname.lastname@kit.edu](mailto:firstname.lastname@kit.edu)

**Abstract**

This paper describes the phrase-based SMT systems developed for our participation in the WMT11 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 and fine-grained POS language models, POS-based reordering, lattice phrase extraction and discriminative word alignment. Furthermore, we present a special filtering method for the English-French Giga corpus and the phrase scoring step in the training is parallelized.

**1 Introduction**

In this paper we describe our systems for the EMNLP 2011 Sixth Workshop on Statistical Ma- chine 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 stan- dard log-linear model combination of phrase-based MT. These include advanced reordering models and corresponding adaptations to the phrase extraction process as well as extension to the translation and language model in form of discriminative word alignment and a bilingual language model to ex- tend source word context. For English-German, lan- guage models based on fine-grained part-of-speech tags were used to address the difficult target lan- guage generation due to the rich morphology of Ger- man.

We also present a filtering method directly ad- dressing the problems of web-crawled corpora, which enabled us to make use of the French-English Giga corpus. Another novelty in our systems this year is the parallel phrase scoring method that re- duces the time needed for training which is espe- cially convenient for such big corpora as the Giga corpus.

**2 System Description**

The baseline systems for all languages use a trans- lation model that is trained on EPPS and the News Commentary corpus and the phrase table is based on a GIZA++ word alignment. The language model was trained on the monolingual parts of the same corpora by the SRILM Toolkit (Stolcke, 2002). It is a 4-gram SRI language model using Kneser-Ney smoothing.

The problem of word reordering is addressed us- ing the POS-based reordering model as described in Section 2.4. The part-of-speech tags for the re- ordering model are obtained using the TreeTagger (Schmid, 1994).

An in-house phrase-based decoder (Vogel, 2003) is used to perform translation and optimization with regard to the BLEU score is done using Minimum Error Rate Training as described in Venugopal et al. (2005). During decoding only the top 20 translation options for every source phrase were considered.

**2.1 Data**

We trained all systems using the parallel EPPS and News Commentary corpora. In addition, the UN corpus and the Giga corpus were used for training

379

*Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 379–385, Edinburgh, Scotland, UK, July 30–31, 2011. *Q*c 2011 Association for Computational Linguistics

the French-English systems.

Optimization was done for most languages using the news-test2008 data set and news-test2010 was used as test set. The only exception is German- English, where news-test2009 was used for opti- mization due to system combination arrangements. The language models for the baseline systems were trained on the monolingual versions of the training corpora. Later on, we used the News Shuffle and the Gigaword corpus to train bigger language models. For training a discriminative word alignment model, a small amount of hand-aligned data was used.

**2.2 Preprocessing**

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

For the German parts of the training corpus we use the hunspell1 lexicon to map words writ- ten according to old German spelling to new Ger- man spelling, to obtain a corpus with homogenous spelling.

Compound splitting as described in Koehn and Knight (2003) is applied to the German part of the corpus for the German-to-English system to reduce the out-of-vocabulary problem for German com- pound words.

**2.3 Special filtering of the Giga parallel Corpus**

The Giga corpus incorporates non-neglegible amounts of noise even after our usual preprocess- ing. This noise may be due to different causes. For instance: non-standard HTML characters, meaningless parts composed of only hypertext codes, sentences which are only partial translation of the source, or eventually not a correct translation at all.

Such noisy pairs potentially degrade the transla- tion model quality, therefore it seemed more conve- nient to eliminate them.

Given the size of the corpus, this task could not be performed manually. Consequently, we used an au- tomatic classifier inspired by the work of Munteanu

sifier should be able to filter out the pairs which likely are not beneficial for the translation model.

In order to reliably decide about the classifier to use, we evaluated several techniques. The training and test sets for this evaluation were built respec- tively from nc-dev2007 and nc-devtest2007. In each set, about 30% randomly selected source sentences switch positions with the immediate following so that they form negative examples. We also used lex- ical dictionaries in both directions based on EPPS and UN corpora.

We relied on seven features in our classifiers: IBM1 score in both directions, number of unaligned source words, the difference in number of words be- tween source and target, the maximum source word fertility, number of unaligned target words, and the maximum target word fertility. It is noteworthy that all the features requiring alignment information (such as the unaligned source words) were computed on the basis of the Viterbi path of the IBM1 align- ment. The following classifiers were used:

**Regression** Choose either class based on a weighted linear combination of the features and a fixed threshold of 0.5.

**Logistic regression** The probability of the class is expressed as a sigmoid of a linear combination of the different features. Then the class with the highest probability is picked.

**Maximum entropy classifier** We used the same set of features to train a maximum entropy classi- fier using the Megam package2.

**Support vector machines classifier** An SVM clas- sifier was trained using the SVM-light pack- age3.

Results of these experiments are summarized in

Table 1.

The regression weights were estimated so that to minimize the squared error. This gave us a pretty poor F-measure score of 90.42%. Given that the lo- gistic regression is more suited for binary classifica- tion in our case than the normal regression, it led to significant increase in the performance. The training

and Marcu (2005) on comparable corpora. This clas-

2 [http://www.cs.utah.edu/˜hal/megam/](http://www.cs.utah.edu/%CB%9Chal/megam/)

1 <http://hunspell.sourceforge.net/>

3 <http://svmlight.joachims.org/>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Approach | Precision | Recall | F-measure | **2.4.2 Lattice Phrase Extraction** |
| Regression | 93.81 | 87.27 | 90.42 | For the test sentences, the POS-based reordering |
| LogReg | 93.43 | 94.84 | 94.13 | allows us to change the word order in the source sen- |
| MaxEnt | 93.69 | 94.54 | 94.11 | tence so that the sentence can be translated more eas- |
| SVM | 98.20 | 96.87 | 97.53 | ily. If we apply this also to the training sentences, we |

Table 1: Results of the filtering experiments

was held by maximizing the likelihood to the data with *L*2 regularization (with *α* = 0*.*1). This gave an F-measure score of 94.78%.

The maximum entropy classifier performed better than the logistic regression in terms of precision but however it had worse F-measure.

Significant improvements could be noticed us- ing the SVM classifier in both precision and recall:

98.20% precision, 96.87% recall, and thus 97.53% F-measure.

As a result, we used the SVM classifier to filter the Giga parallel corpus. The corpus contained orig- inally around 22.52 million pairs. After preprocess- ing and filtering it was reduced to 16.7 million pairs. Thus throwing around 6 million pairs.

**2.4 Word Reordering**

In contrast to modeling the reordering by a distance- based reordering model and/or a lexicalized distor- tion model, we use a different approach that relies on part-of-speech (POS) sequences. By abstracting from surface words to parts-of-speech, we expect to model the reordering more accurately.

**2.4.1 POS-based Reordering Model**

To model reordering we first learn probabilistic rules from the POS tags of the words in the train- ing corpus and the alignment information. Contin- uous reordering rules are extracted as described in Rottmann and Vogel (2007) to model short-range re- orderings. When translating between German and English, we apply a modified reordering model with non-continuous rules to cover also long-range re- orderings (Niehues and Kolss, 2009). The reorder- ing rules are applied to the source text and the orig- inal order of words and the reordered sentence vari- ants generated by the rules are encoded in a word lattice which is used as input to the decoder.

would be able to extract the phrase pairs for orig-

inally discontinuous phrases and could apply them during translation of reordered test sentences.

Therefore, we build reordering lattices for all training sentences and then extract phrase pairs from the monotone source path as well as from the re- ordered paths.

To limit the number of extracted phrase pairs, we extract a source phrase only once per sentence even if it is found in different paths.

**2.5 Translation and Language Models**

In addition to the models used in the baseline sys- tem described above we conducted experiments in- cluding additional models that enhance translation quality by introducing alternative or additional in- formation into the translation or language modelling process.

**2.5.1 Discriminative Word Alignment**

In most of our systems we use the PGIZA++ Toolkit4 to generate alignments between words in the training corpora. The word alignments are gen- erated in both directions and the grow-diag-final-and heuristic is used to combine them. The phrase ex- traction is then done based on this word alignment.

In the English-German system we applied the Discriminative Word Alignment approach as de- scribed in Niehues and Vogel (2008) instead. This alignment model is trained on a small corpus of hand-aligned data and uses the lexical probability as well as the fertilities generated by the PGIZA++ Toolkit and POS information.

**2.5.2 Bilingual Language Model**

In phrase-based systems the source sentence is segmented by the decoder according to the best com- bination 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

4 [http://www.cs.cmu.edu/˜qing/](http://www.cs.cmu.edu/%CB%9Cqing/)

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, an additional lan- guage model in the phrase-based system in which each token consist 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).

**2.5.3 Parallel phrase scoring**

The process of phrase scoring is held in two runs. The objective of the first run is to compute the nec- essary counts and to estimate the scores, all based on the source phrases; while the second run is sim- ilarly held based on the target phrases. Thus, the extracted phrases have to be sorted twice: once by source phrase and once by target phrase. These two sorting operations are almost always done on an ex- ternal storage device and hence consume most of the time spent in this step.

The phrase scoring step was reimplemented in or- der to exploit the available computation resources more efficiently and therefore reduce the process- ing time. It uses optimized sorting algorithms for large data volumes which cannot fit into memory (Vitter, 2008). In its core, our implementation re- lies on STXXL: an extension of the STL library for external memory (Kettner, 2005) and on OpenMP for shared memory parallelization (Chapman et al.,

2007).

Table 2 shows a comparison between Moses and our phrase scoring tools. The comparison was held using sixteen-core 64-bit machines with 128 Gb RAM, where the files are accessed through NFS on a RAID disk. The experiments show that the gain grows linearly with the size of input with an average of 40% of speed up.

**2.5.4 POS Language Models**

In addition to surface word language models, we did experiments with language models based on part-of-speech for English-German. We expect that having additional information in form of probabil- ities of part-of-speech sequences should help espe- cially in case of the rich morphology of German and

#pairs(G) Moses *∗*103(s) KIT *∗*103(s)

0.203 25.99 17.58

1.444 184.19 103.41

1.693 230.97 132.79

Table 2: Comparison of Moses and KIT phrase extraction systems

therefore the more difficult target language genera- tion.

The part-of-speeches were generated using the

TreeTagger and the RFTagger (Schmid and Laws,

2008), which produces more fine-grained tags that include also person, gender and case information. While the TreeTagger assigns 54 different POS tags to the 357K German words in the corpus, the RF- Tagger produces 756 different fine-grained tags on the same corpus.

We tried n-gram lengths of 4 and 7. While no im- provement in translation quality could be achieved using the POS language models based on the normal POS tags, the 4-gram POS language model based on fine-grained tags could improve the translation system by 0.2 BLEU points as shown in Table 3. Surprisingly, increasing the n-gram length to 7 de- creased the translation quality again.

To investigate the impact of context length, we performed an analysis on the outputs of two different systems, one without a POS language model and one with the 4-gram fine-grained POS language model. For each of the translations we calculated the aver- age length of the n-grams in the translation when applying one of the two language models using 4- grams of surface words or parts-of-speech. The re- sults are also shown in Table 3.

The average n-gram length of surface words on the translation generated by the system without POS language model and the one using the 4-gram POS language model stays practically the same. When measuring the n-gram length using the 4-gram POS language model, the context increases to 3.4. This increase of context is not surprising, since with the more general POS tags longer contexts can be matched. Comparing the POS context length for the two translations, we can see that the context in- creases from 3.18 to 3.40 due to longer matching POS sequences. This means that the system using

|  |  |  |  |
| --- | --- | --- | --- |
| the POS language model actually generates trans- | System | Dev | Test |
| lations with more probable POS sequences so that | Baseline | 18.49 | 19.10 |
| longer matches are possible. Also the perplexity | + NewsShuffle LM | 20.63 | 22.24 |
| drops by half since the POS language model helps | + LongRange Reordering | 21.00 | 22.68 |
| constructing sentences that have a better structure. | + Additional Giga LM | 21.80 | 22.92 |

System BLEU avg. ngram length PPL Word POS POS

no POS LM 16.64 2.77 3.18 66.78

POS LM 16.88 2.81 3.40 33.36

Table 3: Analysis of context length

**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 ex- periments 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 German-to-English baseline system applies short-range reordering rules and uses a language model trained on the EPPS and News Commen- tary. By exchanging the baseline language model by one trained on the News Shuffle corpus we im- prove the translation quality considerably, by more than 3 BLEU points. When we expand the cov- erage of the reordering rules to enable long-range reordering we can improve even further by 0.4 and adding a second language model trained on the En- glish Gigaword corpus we gain another 0.3 BLEU points. To ensure that the phrase table also includes reordered phrases, we use lattice phrase extraction and can achieve a small improvement. Finally, a bilingual language model is added to extend the con- text of source language words available for transla- tion, reaching the best score of 23.35 BLEU points. This system was used for generating the translation submitted to the German-English Translation Task.

**3.2 English-German**

The English-to-German baseline system also in- cludes short-range reordering and uses translation

+ Lattice Phrase Extraction 21.87 22.96

+ Bilingual LM **22.05 23.35**

Table 4: Translation results for German-English

and language model based on EPPS and News Com- mentary. Exchanging the language model by the News Shuffle language model again yields a big im- provement by 2.3 BLEU points. Adding long-range reordering improves a lot on the development set while the score on the test set remains practically the same. Replacing the GIZA++ alignments by alignments generated using the Discriminative Word Alignment Model again only leads to a small im- provement. By using the bilingual language model to increase context we can gain 0.1 BLEU points and by adding the part-of-speech language model with rich parts-of-speech including case, number and gender information for German we achieve the best score of 16.88. This system was used to gener- ate the translation used for submission.

|  |  |  |
| --- | --- | --- |
| System | Dev | Test |
| Baseline | 13.55 | 14.19 |
| + NewsShuffle LM | 15.10 | 16.46 |
| + LongRange Reordering | 15.79 | 16.46 |
| + DWA | 15.81 | 16.52 |
| + Bilingual LM | 15.85 | 16.64 |
| + POS LM | **15.88** | **16.88** |

Table 5: Translation results for English-German

**3.3 English-French**

Table 6 summarizes how our system for English- French evolved. The baseline system for this direc- tion was trained on the EPPS and News Commen- tary corpora, while the language model was trained on the French part of the EPPS, News Commen- tary and UN parallel corpora. Some improvement could be already seen by introducing the short-range reorderings trained on the baseline parallel corpus.

Apparently, the UN data brought only slight im- provement to the overall performance. On the other

model adding around 0.2 to the previous score and thus giving 28.34 as final score.

hand, adding bigger language models trained on the

monolingual French version of EPPS, News Com- mentary and the News Shuffle together with the French Gigaword corpus introduces an improvement of 3.7 on test. Using a system trained only on the Giga corpus data with the same last configuration shows a significant gain. It showed an improvement of around 1.0. We were able to obtain some further improvements by merging the translation models of the last two systems. i.e. the one system based on EPPS, UN, and News Commentary and the other on the Giga corpus. This merging increased our score by 0.2. Finally, our submitted system for this direc- tion was obtained by using a single language model trained on the union of all the French corpora in- stead of using multiple models. This resulted in an improvement of 0.1 leading to our best score: 28.28.

|  |  |  |
| --- | --- | --- |
| System | Dev | Test |
| Baseline | 20.62 | 22.36 |
| + Reordering | 21.29 | 23.11 |
| + UN | 21.27 | 23.24 |
| + Big LMs | 23.77 | 26.90 |
| Giga data | 24.53 | 27.94 |
| Merge | 24.74 | 28.14 |
| + Merged LMs | **25.07** | **28.28** |

Table 6: Translation results for English-French

**3.4 French-English**

The development of our system for the French- English direction is summarized in Table 7. Our sys- tem for this direction evolved quite similarly to the opposite direction. The largest improvement accom- panied the integration of the bigger language mod- els (trained on the English version of EPPS, News Commentary, News Shuffle and the Gigaword cor- pus): 3.3 BLEU points, whereas smaller improve- ments could be gained by applying the short reorder- ing rules and almost no change by including the UN data. Further gains were obtained by training the system on the Giga corpus added to the previous parallel data. This increased our performance by

0.6. The submitted system was obtained by aug- menting the last system with a bilingual language

System Dev Test

Baseline 20.76 23.78

+ Reordering 21.42 24.28

+ UN 21.55 24.21

+ Big LMs 24.16 27.55

+ Giga data 24.86 28.17

+ BiLM **25.01 28.34**

Table 7: Translation results for French-English

**4 Conclusions**

We have presented the systems for our participation in the WMT 2011 Evaluation for English*↔*German and English*↔*French. For English*↔*French, a spe-

cial filtering method for web-crawled data was de- veloped. In addition, a parallel phrase scoring tech- nique was implemented that could speed up the MT training process tremendously. Using these two fea- tures, we were able to integrate the huge amounts of data available in the Giga corpus into our systems translating between English and French.

We applied POS-based reordering to improve our translations in all directions, using short-range re-

ordering for English*↔*French and long-range re-

ordering for English*↔*German. For German-

English, reordering also the training corpus lead to

further improvements of the translation quality.

A Discriminative Word Alignment Model led to an increase in BLEU for English-German. For this direction we also tried fine-grained POS language models of different n-gram lengths. The best trans- lations could be obtained by using 4-grams.

For nearly all experiments, a bilingual language model was applied that expands the context of source words that can be considered during decod- ing. The improvements range from 0.1 to 0.4 in BLEU score.

**Acknowledgments**

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

**References**

Barbara Chapman, Gabriele Jost, and Ruud van der Pas.

2007. *Using OpenMP: Portable Shared Memory Par- allel Programming (Scientific and Engineering Com- putation)*. The MIT Press.

Roman Dementiev Lutz Kettner. 2005. Stxxl: Stan- dard template library for xxl data sets. In *Proceedings of ESA 2005. Volume 3669 of LNCS*, pages 640–651. Springer.

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

Dragos Stefan Munteanu and Daniel Marcu. 2005. Im- proving machine translation performance by exploit- ing non-parallel corpora. *Computational Linguistics*,

31:477–504.

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, Stephan Vogel, and Alex Waibel. 2011. Wider Context by Using Bilingual Lan- guage Models in Machine Translation. In *Sixth Work- shop on Statistical Machine Translation (WMT 2011)*, Edinburgh, UK.

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.

Kay Rottmann and Stephan Vogel. 2007. Word Reorder- ing in Statistical Machine Translation with a POS- Based Distortion Model. In *TMI*, Sko¨ vde, Sweden.

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

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

Andreas Stolcke. 2002. SRILM – An Extensible Lan- guage 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 *Work- shop on Data-drive Machine Translation and Beyond (WPT-05)*, Ann Arbor, MI.

Jeffrey Scott Vitter. 2008. *Algorithms and Data Struc- tures for External Memory*. now Publishers Inc.

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