# The KIT Translation system for IWSLT 2010

*Jan Niehues*[1], *Mohammed Mediani*[1], *Teresa Herrmann*[1], *Michael Heck*[2], *Christian Herff*[2], *Alex Waibel*[1]

Institute of Anthropomatics
KIT - Karlsruhe Institute of Technology
[1]firstname.lastname@kit.edu
[2]firstname.lastname@student.kit.edu

## Abstract

This paper presents the KIT systems participating in the French to English BTEC and in the English to French TALK Translation tasks in the framework of the IWSLT 2010 machine translation evaluation.

Starting with a state-of-the art phrase-based translation system we tested different modifications and extensions to improve the translation quality of the system.

First, we improved the word reordering by learning POS-based reordering rules from an automatically word-aligned parallel corpus. Furthermore, different experiments to adapt the machine translation system towards the target domain were carried out. In addition, for the BTEC task we tried to avoid data-sparseness problems by using word stems instead of the full word forms.

## 1. Introduction

In this paper we describe the systems that we built for our participation in the IWSLT TALK task and the IWSLT BTEC French to English task. We used a state-of-the-art phrase-based machine translation system to generate the translations for both tasks. We extended the different models of the baseline system to be able to improve the translation performance on both systems.

One of the biggest problems in machine translation is the different word order between languages. Therefore, we applied the POS-based reordering model presented in [1] which has already been successfully used in other text translation evaluations [2] to improve the word order in the generated target sentence.

Whereas a considerable amount of out-of-domain data was available for the TALK task, the in-domain data provided by the evaluation was very limited. Therefore, we applied different methods to adapt the models of the translation system to the target domain. Furthermore, we extended the translation model by a bilingual language model.

For the BTEC task only a very small set of training sentences could be used. Consequently, we tried to handle the data-sparseness problems by using the word stems instead of the full word forms.

The following sections evolve in a similar way as the system development. First, a short description of the different training data sets is given. Then the preprocessing and postprocessing techniques are briefly discussed and the baseline system is presented. Sections presenting different special enhancements will follow after that, starting with the POS-based reordering model and the translation model adaptation followed by bilingual language models and stemming. Finally, a summary of the whole process is given in numbers. We close our paper with the conclusions drawn from this work and some future perspectives.

### 1.1. Training data

In addition to the data provided for the TALK task (TED talks), we used parallel data from different open domain sources. This includes: United Nations documents (UN), European Parliament Proceedings (EPPS) and the News Commentary (NC) corpus. For language modeling, we used all the target parts of the parallel data. Moreover, we used the French part of the Giga parallel corpus provided for the evaluation.

Since only one small tuning data set was made officially available, we chose to use it as a test set. In order to tune the system parameters, we split apart a small subset from the provided training data consisting of the most recent talks of the TED parallel corpus with a size comparable to the test set (i.e. the set originally provided for tuning).

Table 1 summarizes data statistics.

Table 1: Brief Statistics of the Evaluation Data

| | | **English** | **French** |
|---|---|---|---|
| Parallel | TED | 702.69K | 730.88K |
| | UN | 198.11M | 227.35M |
| | EPPS | 39.17M | 43.02M |
| | NC | 1.81M | 2.09M |
| | BTEC | 208.44K | 217.32K |
| TALK Dev | | 8.22K | 8.82K |
| BTEC Dev | | 68.65K | 4.53K |
| TALK Test | | 10.95K | 10.74K |
| BTEC Test | | 55.02K | 3.16K |
| Monolingual | | - | 672.07M |

## 2. Preprocessing and postprocessing

All the training data was preprocessed before the models were trained. In this step different normalizations were applied such as mapping different types of quotes and normalizing dates and numbers. After that, the first word of every sentence was smartcased. Finally, we removed sentences that are too long and empty lines to obtain the final training corpus.

A special preprocessing procedure was applied to the large Giga corpus since it is an extremely noisy source. We first built a lexicon from the EPPS corpus and used it to lexically score different pairs in the noisy corpus in a manner similar to the IBM Model 1. A threshold was manually set to filter out the worst scoring pairs.

The postprocessing is only about recasing. Given the nature of TED talks, where every line might be an incomplete sentence, we were more careful about casing the first letter of a sentence. Indeed, we only recased the first letter if it follows a sentence terminated by an end-of-sentence mark (a period for instance).

## 3. Baseline System

For both tasks we used a phrase-based Statistical Machine Translation (SMT) system as a baseline. The system was trained on the aforementioned monolingual and parallel data.

To build the translation model, a word alignment of the parallel corpus was generated by the GIZA++-Toolkit [3] for both directions. Afterwards, the alignments were combined using the grow-diag-final-and heuristic. Then the translation model was built using the Moses Toolkit [4]. In the TALK task, we used only the EPPS, NC and TED corpus as parallel data for the baseline system.

The language models were trained using the SRI Language Modeling Toolkit [5]. For all systems, we used a 4-gram language model and applied modified Kneser-Ney smoothing [6]. For the TALK task, the language model for the baseline system was built by a linear interpolation of three SRI language models based on the target parts of the UN, EPPS, and NC parallel data. The weights for the interpolation were selected in a way that they minimize the perplexity on the development data.

In order to generate the translations we used our phrase-based, in-house decoder [7]. In the baseline system, we used a distance-based reordering model. To find the best performing weights of the different features in the log-linear model we used Minimum Error Rate Training as proposed in [8]. We used the BLEU score as an error metric in the optimization.

## 4. Word Reordering Model

One part of our system that differs from the baseline system is the reordering model. To account for different word order in the languages, we used the POS-based reordering model presented in [1]. This model learns rules from a parallel text to reorder the source side. The goal is to generate a reordered source side that can be translated in a more monotone way.

In this framework, POS information was added to the source side. The POS tags were generated by the TreeTagger [9]. It uses a tag set of 57 tags. Afterwards, reordered rules of length up to 10 tags were extracted from the aligned parallel corpus. The alignment was generated by combining the GIZA++ alignments of both directions using the grow-diag-final-and heuristic. These rules are of the form *VVIMP VMFIN PPER* $\rightarrow$ *PPER VMFIN VVIMP* and describe how the source side has to be reordered to match the target side. Then the rules are pruned and the remaining ones are scored according to their relative frequencies.

The learned rules are then applied to the test sentences in a preprocessing step to the actual decoding. Therefore, first the POS information of the test data is generated, also by the TreeTagger. Then different reorderings of the source sentences that better match the word order in the target language are encoded in a word lattice. First, the lattice consists only of the monotone path. Afterwards, we search for reordering rules that can be applied to a sentence. For all of them, a path is added to the lattice. The edges of the path are weighted according to the relative frequencies of the rules.

After this preprocessing, the decoding is then performed on the resulting word lattice.

## 5. Adaptation

Although a huge amount of out-of-domain training data was provided, the amount of in-domain training data was much smaller. On the one hand, we want to facilitate the better estimation of the probabilities available through the large amounts of data instead of using only the small in-domain part. On the other hand, we do not want to dilute the domain knowledge encoded in the in-domain part. Therefore, we tried to adapt the huge models extracted from the out-of-domain data towards the target domain.

Adaptation was performed on two models: on the translation models and on the language models.

### 5.1. Translation model adaptation

The adaptation in this case consists of a combination of two translation models. The bigger model is built from all the available data including the in-domain part. The smaller translation model is trained only on the in-domain data. Since the the alignment does not strongly depend on the domain, we extract it from the complete corpus. Then we proceed to model combination as described hereafter.

A phrase pair with features $\alpha$ from the first model is added to the combination with features $< \alpha, \beta >$, where $\beta$ is a vector of two additional scores. $\beta$ consists of the smoothed relative frequencies of both directions from the smaller model if the phrase pair occurs in it, otherwise it is set to the worst score in the model.

### 5.2. Language model adaptation

On the other side, language model adaptation was realized by using a small in-domain language model in addition to the main language model. Both language models were used in the log-linear model combination. The weights for both models were found during MERT. They were selected in a way that the BLEU score of the whole system is maximized on the development data.

## 6. Bilingual language models

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 [10], 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 French sentence *Je suis rentré à la maison* with the English translation *I went home*, the resulting bilingual text would look like this: *I_Je went_suis_rentré home_la_maison*.

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, we only used a language model based on the target word order.

## 7. Stemming

When translating from French to English, most of the morphological information in the French source text is not necessarily needed in order to generate the English target text, since English is a less inflectional language. The richer morphology of French might lead to sparse data problems, i.e. we might encounter a word in the text to be translated that has not been seen in the training data in the same inflectional variant, but maybe exists with a different inflectional suffix. Especially for adjectives, it does not make any difference whether we are translating the feminine or the masculine instance of an adjective into English, since it would be realized in the same surface form.

We could therefore argue that it might help in the translation process to ignore the morphological inflections on the French side and use only the stems of the French words when translating into English. In order to test this hypothesis, we executed experiments stemming the French source part of the training corpus and trained a system on the stemmed French and fully inflected English data. For the translation step, we then also stemmed the French input text.

## 8. Results

The development of our submitted systems for the English-French TALK translation task and for the French-English BTEC Translation task is summarized in the following. Their performance is measured applying the BLEU metric.

### 8.1. TALK Task

Here we describe the steps and techniques followed to construct our English-French system participating in the TALK task.

As described in Section 3, the baseline for this task was trained on the EPPS, NC and TED parallel data with a reordering window of two words and a 4-gram language model linearly interpolated from the target parts of the parallel corpus plus the UN French part. With this configuration a score around 27 on the development set and around 23 on the test set was obtained.

The reordering model we used in the baseline system does not model the reordering problem accurately. It does not distinguish which words take part in the reordering, but only the length of the reordering. To include more information about the word's grammatical categories we used the POS-based reordering as described in Section 4. By using this approach we gained around 0.3 on the development set and 0.8 on the test set.

The adaptation idea seems more attractive when the domain of interest is somehow different compared to the greatest amount of available data. This is the reason why our next couple of experiments concentrated on this idea. First, using the in-domain language model constructed only from the TED data together with the main out-of-domain one gives an important improvement on dev, more than 1 BLEU point, and around 0.7 on test. After that, we got more improvement of around 0.6 BLEU points on the development set and around 1 on the test set by adapting the phrase table to the target domain. Therefore, we built a small phrase table based on the TED data only and combined it with the big phrase table as described above.

More parallel data was available in the UN corpus. Exploiting it was not that helpful on the development set, it decreases the score on this set by 0.17 whereas it increased the test set score by 0.1.

In the next experiment, we applied a compound adaptation in two steps to the big translation model trained on all the available parallel data (i.e. UN, EPPS, NC, and TED). First, the adaptation was performed with the translation model built from the EPPS, NC, and TED parallel data. The resulting model is next adapted a second time to the smaller TED translation model. This approach compensated the loss faced by introducing the UN parallel data and added an extra 0.1

BLEU points on the development set, while the score on the test set remained almost unchanged.

Table 2 shows example translations obtained from the system without 2-step adaptation (System 5 in Table 4) and the system applying the 2-step adaptation (System 6) together with the corresponding source and reference text. Here we can see how applying this type of adaptation helps to discount the very general content stemming from the UN corpus and to prioritize the data that is more relevant for the TALK task. In this particular case the adapted system is able to disambiguate the two senses of the English *grave* and chooses the correct French word in this context, i.e. *tombe*.

Table 2: Example Translation without (-) and with (+) 2-step Adaptation

| | |
|---|---|
| src | ... and carries with him into the **grave** the last syllables ... |
| - | ... et la transporte avec lui dans la **gravité** de la derniére syllabes ... |
| + | ... et la transporte avec lui dans la **tombe** la demiére syllabes ... |
| ref | ... et emporte les derniéres syllabes avec lui dans la **tombe** ... |

After exploiting possible adaptations, our next experiment was about the effect of a bilingual language model on the system. In fact, in addition to the language models used so far, we introduced another one to the decoder, that is the bilingual model built as explained in Section 6 from the parallel corpus used in the first experiments (i.e. UN, EPPS, NC, and TED). Important gains were observed in this experiment: almost 0.5 on the development set and 0.3 on the test set.

Table 3 presents translations from the system without bilingual language model (System 6) and with bilingual language model (System 7). This example shows how introducing the bilingual language model affects the translation in a positive way, here by correcting the grammatical number of the verb. Using the information about the aligned source words available in the bilingual language model, the decoder is able to choose the singular verb form *parle* over the plural form *parlent* as a translation in this sentence.

Table 3: Example Translation without (-) and with (+) Bilingual Language Model

| | |
|---|---|
| src | ... you must marry **someone who speaks** a different language. |
| - | ... vous devez épouser **quelqu'un qui parlent** une langue différente. |
| + | ... vous devez épouser **quelqu'un qui parle** une langue différente. |
| ref | ... vous devez épouser **une personne qui parle** une langue différente. |

In our last experiment we added an extra language model to the system constructed from the Giga corpus. The test set was affected much more by this additional model than the development set. Indeed, a gain of 1.3 BLEU points on the test set was obtained whereas a gain of 0.3 was achieved on the development set. This led to our final system which was used to generate the submission with a final score of 30.39 on the development set and 26.34 on the test set. Table 4 summarizes the development steps together with the corresponding scores.

Table 4: Translation Results for English-French TALK Translation Task (BLEU scores)

| | System | Dev | Test |
|---|---|---|---|
| 1 | Baseline | 27.39 | 22.70 |
| 2 | + Short-range Reordering | 27.72 | 23.56 |
| 3 | + Language Model Adaptation | 28.94 | 24.29 |
| 4 | + Translation Model Adaptation | 29.57 | 24.68 |
| 5 | + UN corpus | 29.40 | 24.78 |
| 6 | + 2-step Adaptation | 29.62 | 24.67 |
| 7 | + Bilingual Language Model | 30.08 | 24.95 |
| 8 | + Giga Language Model | 30.39 | 26.34 |

### 8.2. TALK Task ASR Output

In the TALK task of the IWSLT 2010 Evaluation not only the reference transcript of the TED talks, but also the automatic transcription of the talks were to be translated. Since there are additional errors in the automatic transcription, we experimented with different approaches to handle these errors. We carried out some experiments on handling errors in casing as well as punctuation errors on the source side. The experiments were performed using two different systems. The first system we used is the one described as System 4 in Table 4. The other one is the final system also described in that table. The results for the experiments on ASR output are shown in Table 5. The columns indicate different scoring variants, where C1 represents scoring case-sensitively and with punctuation, C2 denotes case-sensitive scoring without punctuation and C3 case-insensitive scoring without punctuation.

Since the ASR output is not always cased correctly, we tried to ignore this information on the source side and generate the case anew for the target side. Therefore, we lowercase the ASR text as well as the source side of the phrase table. This generates a translation system translating from lower-cased English to mixed-case French. As shown in the table, the performance of this approach is similar to the baseline one of the first system, but worse when applied to the final system.

The second approach tries to improve the punctuation on the target side. We tried the same approach as for the casing, by translating source text without punctuation into punctuated target text. Therefore, we removed the punctuation

from the input text and from the source side of the phrase table. For the first system we could improve the performance when scored with punctuation, but the performance decreased when punctuation is ignored in scoring. We could further improve this system by also ignoring the case information.

In contrast, if we performed this experiment on the final system, we could not improve the performance of the system as shown in the lower part of Table 5. Therefore, we applied no special treatment of case and punctuation for translating the ASR outputs.

Table 5: Translation Results for English-French TALK Translation Task ASR Output (BLEU Scores)

| System | C1 | C2 | C3 |
|---|---|---|---|
| No UN System | 13.45 | 14.65 | 16.05 |
| Ignore Case | 13.42 | 14.58 | 16.13 |
| Ignore Punc | 13.86 | 14.15 | 15.67 |
| Ignore Case & Punc | 14.06 | 14.24 | 15.82 |
| Final System | 14.47 | 15.65 | 17.13 |
| Ignore Case | 14.13 | 15.45 | 16.93 |
| Ignore Punc | 14.29 | 15.13 | 16.62 |
| Ignore Case & Punc | 13.85 | 14.76 | 16.24 |

### 8.3. BTEC Task

The French to English translation system for the BTEC task was trained on the training corpus and the CStar development set provided for the task. As development set the IWSLT'04 set was used and for testing the sentences from the IWSLT'05 were translated. During decoding a reordering window of length 4 was used. We used a 4-gram language model trained on the training corpus as well as on the CStar development set using the SRI Language Modeling Toolkit.

The baseline system achieved a performance of 59.17 BLEU points on the test set. For the next system we replace the words by their stems in the French source text. The stems are obtained using the TreeTagger [9]. Since the different inflections of a French word are often translated to the same English words, we hardly lose any information.

But especially for a small corpus like the one used in this task, the stemming leads to a better estimation of the probabilities. Furthermore, more words can be translated, since all word forms of a stem can be translated even though only some may have been seen during training. This approach could improve the translation performance by 1.6 BLEU points up to 60.83.

In the next experiment we evaluated whether the gain is only due to a better estimated alignment. Therefore, we used the alignment generated on the stemmed corpus to extract phrase pairs from the original corpus. This leads to a BLEU score of 60.73. So it seems that most of the improvement comes from the better alignment, but the best result is achieved by using it also in the phrase extraction.

Afterwards, we tested the POS-based word reordering on the best performing system. In contrast to previous experiments on different tasks, for this task we could not improve the performance. The reason may be, that the sentences are quite short and the word order in French and English is quite similar, so word reordering seems not to be the big problem.

In a last step, we improved our postprocessing by replacing *are*, *am* and *would* by their contracted forms. This brought about an improvement of additional 0.7 BLEU points leading to the best system that could achieve a performance of 61.51 BLEU points.

Table 6: Translation Results for French-English BTEC Task (BLEU score)

| | System | Test |
|---|---|---|
| 1 | Baseline | 59.17 |
| 2 | 1 + source stems | 60.83 |
| 3 | 1 + only alignment from stemmed source | 60.73 |
| 4 | 2 + POS-based Reordering | 60.55 |
| 5 | 2 + add. Postprocessing | 61.51 |

### 8.4. Conclusions

In this paper, we presented the systems we built for generating our submission to the English-French TALK Translation and French-English BTEC Translation tasks for the IWSLT 2010 Evaluation. The development of both systems is heavily based on the STTK decoder, even though the Moses Toolkit was also used to align words and extract phrases.

In the TALK Translation system, we dealt with new issues such as incomplete sentences for which we had to perform different procedures. For instance, casing the first letter was only applied if the previous segment ended with a sentence termination mark.

Even though the reordering distances between French and English are not that far, using the POS reordering rules noticably improved the translation quality in the TALK task. By contrast, in the BTEC task a 4-word window was apparently enough to gather all the reordering information and therefore the POS reordering had no effect.

From another side, given that the talks are somehow different from the usual parallel training data and since only a very limited amount of talk data was available, it seems that the adaptation has a positive effect on the translation quality. Due to this, we needed an additional adaptation step in order to compensate for the bias introduced by more parallel data from the general domain.

Apparently, optimizing a weighted linear combination of multiple language models with the STTK decoder was an efficient way to exploit more than one model. Alternatively, merging the data and initiating a new training process would be time and resource consuming. Moreover, the bilingual

language model extracted from parallel data showed interesting improvements.

It has been shown that morphological analyses are of great help when translating between morphologically rich and less rich languages [11]. This was exactly the case for the BTEC task. Definitely, the most useful technique was the stemming. It could absolutely reduce the data sparseness due to the small parallel corpus.

One of the improvements we are thinking of is to apply the bilingual language model approach using fine-grained part-of-speech information in order to facilitate matching congruency, for example between subject and verb or within noun phrases. Moreover, as explained in Section 6 we only used a bilingual language model based on the target side of the parallel data. We expect more success by combining both bilingual language models based on both directions.

Based on the BTEC experiments, it would be interesting to see the effect of including more linguistic clues into the translation especially when the languages are structurally very different.

## 9. Acknowledgements

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

## 10. References

[1] K. Rottmann and S. Vogel, "Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model," in *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Skövde, Sweden, 2007.

[2] J. Niehues, T. Herrmann, M. Mediani, and A. Waibel, "The Karlsruhe Institute of Technology Translation System for the ACL-WMT 2010," in *Proceedings of the Fifth Workshop on Statistical Machine Translation (WMT 2010)*, Uppsala, Sweden, 2010.

[3] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[4] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of ACL 2007, Demonstration Session*, Prague, Czech Republic, 2007.

[5] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit." in *International Conference on Spoken Language Processing*, Denver, Colorado, USA, 2002.

[6] S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling." Computer Science Group, Harvard University, Tech. Rep. TR-10-98, 1998.

[7] S. Vogel, "SMT Decoder Dissected: Word Reordering." in *International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China, 2003.

[8] A. Venugopal, A. Zollman, and A. Waibel, "Training and Evaluation Error Minimization Rules for Statistical Machine Translation," in *Workshop on Data-drive Machine Translation and Beyond (WPT-05)*, Ann Arbor, Michigan, USA, 2005.

[9] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," in *International Conference on New Methods in Language Processing*, Manchester, United Kingdom, 1994.

[10] A. Allauzen, J. Crego, A. Max, and F. Yvon, "LIMSI's statistical translation system for WMT'09," in *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT 2009)*, Athens, Greece, 2009.

[11] Y.-S. Lee, "Morphological analysis for statistical machine translation," in *Proceedings of the Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Boston, Massachusetts, USA, 2004.