The KIT Translation Systems for IWSLT 2012

Mohammed Mediani[⋆]*, Yuqi Zhang*[⋆]*, Thanh-Le Ha*[⋆]*, Jan Niehues*[⋆]*, Eunah Cho*[⋆]*, Teresa Herrmann*[⋆]*, Rainer Kärgel*[†] *and Alexander Waibel*[⋆]

Institute of Anthropomatics
KIT - Karlsruhe Institute of Technology

[⋆] firstname.lastname@kit.edu

[†] rainer.kaergel@student.kit.edu

Abstract

In this paper, we present the KIT systems participating in the English-French TED Translation tasks in the framework of the IWSLT 2012 machine translation evaluation. We also present several additional experiments on the English-German, English-Chinese and English-Arabic translation pairs.

Our system is a phrase-based statistical machine translation system, extended with many additional models which were proven to enhance the translation quality. For instance, it uses the part-of-speech (POS)-based reordering, translation and language model adaptation, bilingual language model, word-cluster language model, discriminative word lexica (DWL), and continuous space language model.

In addition to this, the system incorporates special steps in the preprocessing and in the post-processing step. In the preprocessing the noisy corpora are filtered by removing the noisy sentence pairs, whereas in the postprocessing the agreement between a noun and its surrounding words in the French translation is corrected based on POS tags with morphological information.

Our system deals with speech transcription input by removing case information and punctuation except periods from the text translation model.

1. Introduction

In the IWSLT 2012 Evaluation campaign [1], we participated in the tasks for text and speech translation for the English-French language pair. The TED tasks consist of automatic translation of both the manual transcripts and transcripts generated by automatic speech recognizers for talks held at the TED conferences [1].

The TED talks are given in English in a large number of different domains. Some of these talks are manually transcribed and translated by volunteers over the globe [2]. Given these manual transcripts and a large amount of out-of-domain data (mainly news), our ambition is to perform optimal translation on the untranslated lectures which are more likely from different domains. Furthermore, we strive for performing as well as possible on the automatically transcribed lectures.

The contribution of this work is twofold: on the one hand, it demonstrates how the complementary manipulation of in-domain and out-of-domain data is gainful in building more accurate translation models. It will be shown that while the large amount of out-of-domain data ensures wider coverage, the limited in-domain data indeed helps to model better the style and the genre. On the other hand, we show that using a text translation system with a proper processing of punctuation can handle the translation of automatic transcriptions to some extent.

Compared to our last year's system, three new components are introduced: adaptation of the candidate selection in the translation model (Section 5), continuous space language model (Section 8), and part-of-speech (POS)-based agreement correction (Section 9).

The next section briefly describes our baseline, while Sections 3 through 9 present the different components and extentions used by our phrase-based translation system. These include the special preprocessing of the spoken language translation (SLT) system, POS-based reordering, translation and language model adaptation, the cluster language model, the descriminative word lexica (DWL), the continuous space language model, and the POS-based agreement correction. After that, the results of the different experiments (official and additional language pair systems) are presented and finally a conclusion ends the paper.

2. Baseline System

For the corresponding tasks, the provided parallel data consist of the EPPS, NC, UN, TED and Giga corpora, whereas the monolingual data consist of the monolingual version of the News Commentary and the News Shuffled corpora. In addition, the use of the Google Books Ngrams[2] was allowed. We did not use the UN data and Google Books Ngrams this year. The reason was that in several previous experiments (not reported in this paper), they consistently had a negative impact on the performance.

[1] http://www.ted.com

[2] http://ngrams.googlelabs.com/datasets

A common preprocessing is applied to the raw data before performing any model training. This includes removing long sentences and sentences with length difference exceeding a certain threshold. In addition, special symbols, dates and numbers are normalized. The first letter of every sentence is smart-cased. Furthermore, an SVM classifier was used to filter out the noisy sentences pairs in the Giga English-French corpus as described in [3].

The baseline system was trained on the EPPS, TED, and NC corpora. In addition to the French side of these corpora, we used the provided monolingual data and the French side of the parallel Giga corpus, for language model training. Systems were tuned and tested against the provided Dev 2010 and Test 2010 sets.

All language models used are 4-gram language models with modified Kneser-Ney smoothing, trained with the SRILM toolkit [4]. The word alignment of the parallel corpora was generated using the GIZA++ Toolkit [5] for both directions. Afterwards, the alignments were combined using the grow-diag-final-and heuristic. The phrases were extracted using the Moses toolkit [6] and then scored by our in-house parallel phrase scorer [7]. Phrase pair probabilities are computed using modified Kneser-Ney smoothing as in [8]. Word reordering is addressed using the POS-based reordering model and is described in detail in Section 4. The POS tags for the reordering model are obtained using the TreeTagger [9]. Tuning is performed using Minimum Error Rate Training (MERT) against the BLEU score as described in [10]. All translations are generated using our in-house phrase-based decoder [11].

3. Preprocessing for Speech Translation

The system translating automatic transcripts needs some special preprocessing on the data, since generally there is no or not reliable case information and punctuation in the automatically generated transcripts. We have tried two ways to deal with the difference on casing and punctuation between a machine translation (MT) system and a SLT system. In addition, we also optimize the system with different development data: simulated ASR output and original automatic speech recognition (ASR) output.

In order to make the system translate the automatically generated transcripts, the first method we have used is to lowercase the source side of the training corpora and remove the punctuation except periods from the source language. On these modified source sentences and untouched target sentences, all models are re-trained, including alignments, phrase tables, reordering rules, bilingual language model and DWL model. Therefore, we can avoid having to build a whole MT system for the SLT task. In order to simplify the procedure, we tried a second method where we directly modify the source phrases in the phrase tables. We lowercase the source phrases and remove the punctuation except periods from the source phrases. Though there could be duplicated phrase pairs with different scores in the phrase table due to this modification, during the decoding the phrase with the best scores will be selected according to the weights.

Two ways to optimize the system are possible. The first one is to use the manual transcripts but it requires lower casing and removal of punctuation marks. The other one is to use the ASR single-best output released by the SLT task. The advantage of optimizing with the manual transcripts is that the system will be adjusted with higher quality sentences. On the other side, optimization using ASR output makes the system more consistent with the evaluation test data. We have tested both methods in our experiments.

4. Word Reordering Model

Our word reordering model relies on POS tags as introduced by [12]. Rule extraction is based on two types of input: the Giza alignment of the parallel corpus and its corresponding POS tags generated by the TreeTagger for the source side.

For each sequence of POS tags, where a reordering between source and target sentences is detected, a rule is generated. Its head consists of sequential source tags and its body is the permuted POS tags of the head which match the order of the corresponding aligned target words. After that, the rules are scored according to their occurrence and pruned according to a given threshold.

In our system, the reordering is performed as a preprocessing step. Rules are applied to the test set and possible reorderings are encoded in a word lattice, where the edges are weighted according to the rule's probability.

Finally, the decoding is performed on the resulted word lattice. During decoding, the distance-based phrase reordering could also be applied additionally.

5. Adaptation

To achieve the best performance on the target domain, we performed adaptation for translation models as well as language models.

5.1. Translation Model Adaptation

In a phrase-based translation system, building the translation consists of two steps. First, we select a set of candidate translations from the phrase table (candidate selection). In our system, we normally take the top 10 translations for every source phrase according to initially predefined weights. In the second step, the best translation is built from these candidates using the scores from the translation model (phrase scoring) as well as other models.

In some of our systems we also adapted the first step, while the second step was adapted in all of our systems by using additional scores for the phrase table.

To adapt the translation model towards the target domain, first, a large translation model is trained on all the available data. Then, a separate in-domain model is trained on the in-domain data only, reusing the alignment from the large model. The alignment is trained on the large data, because it

seems to be more important for the alignment to be trained on bigger corpora than being based on only in-domain data.

When we do not adapt the candidate selection, the best translations from the general phrase table is used and only the scores from the in-domain phrase table are taken into account. In the other case, we take the union of the phrase pairs collected from both phrase tables. We will refer to this adaptation method as **CSUnion** in the description of the results.

The scores of the translation model are adapted to the target domain by combining the in-domain and out-of-domain scores in a log-linear combination. The adapted translation model uses the four scores (phrase-pair probabilities and lexical scores for both directions) from the general model as well as the two probabilities of both directions from the small in-domain model. If the phrase pair does not occur in the in-domain part, a default score is used instead of a relative frequency. In our case, we use the lowest probability that occurs in the phrase table.

5.2. Language Model Adaptation

For the language model, it is also important to perform an adaptation towards the target domain. There are several word sequences, which are quite uncommon in general, but may be used often in the target domain.

As it is done for the translation model, the adaptation of the language model is also achieved by a log-linear combination of different models. This also fits well into the global log-linear model used in the translation system. Therefore, we train a separate language model using only the in-domain data from the TED corpus. Then it is used as an additional language model during decoding. Optimal weights are set during tuning by MERT.

6. Cluster Language Model

In addition to the word-based language model, we also use a cluster language model in the log-linear combination. The motivation is to make use of larger context information, since there is less data sparsity when we substitute words by word classes.

First, we cluster the words in the corpus using the MKCLS algorithm [13] given a number of classes. Second, we replace the words in the corpus by their cluster IDs. Finally, we train an n-gram language model on this corpus consisting of cluster IDs.

Because the TED corpus is small and important for this translation task and it exactly matches the target genre, we trained the cluster language model only on TED corpus in our experiments. The TED corpus is characterized by a huge variety of topics, but the style of the different talks of the corpus is quite similar. When translating a new talk from the same domain, we may not find a good translation in the TED corpus for many topic specific words. What TED corpus can help with, however, is to generate sentences in the same style. During decoding the cluster-based language model works as an additional model in the log-linear combination.

7. Discriminative Word Lexica

Mauser et al. [14] have shown that the use of DWL can improve the translation quality. For every target word, they trained a maximum entropy model to determine whether this target word should be in the translated sentence or not using one feature per one source word.

One specialty of this task is that we have a lot of parallel data we can train our models on, but only a quite small portion of these data, the TED corpus, is very important to the translation quality. Since building the classifiers on the whole corpus is quite time consuming, we try to train them on the TED corpus only.

When applying DWL in our experiments, we would like to have the same conditions for the training and test case. For this we would need to change the score of the feature only if a new word is added to the hypothesis. If a word is added the second time, we do not want to change the feature value. In order to keep track of this, additional bookkeeping would be required. Also the other models in our translation system will prevent us from using a word too often.

Therefore, we ignore this problem and can calculate the score for every phrase pair before starting with the translation. This leads to the following definition of the model:

$$p(e|f) = \prod_{j=1}^{J} p(e_j|f) \tag{1}$$

In this definition, $p(e_j|f)$ is calculated using a maximum likelihood classifier.

Each classifier is trained independently on the parallel training data. All sentence pairs where the target word e_j occurs in the target sentence are used as positive examples. We could now use all other sentences as negative examples. But in many of these sentences, we would anyway not generate the target word, since there is no phrase pair that translates any of the source words into the target word.

Therefore, we build a target vocabulary for every training sentence. This vocabulary consists of all target side words of phrase pairs matching a source phrase in the source part of the training sentence. Then we use all sentence pairs where e_j is in the target vocabulary but not in the target sentences as negative examples. This has shown to have a postive influence on the translation quality [3] and also reduces training time.

8. Continuous Space Language Model

In recent years, different approaches to integrate a continuous space models have shown significant improvements in the translation quality of machine translation systems, e.g. [15]. Since the long training time is the main disadvantage of this model, we only trained it on the small, but very domain-relevant TED corpus.

In contrast to most other approaches, we did not use a feed-forward neural network, but used a Restricted Bolzmann Machine (RBM). The main advantage of this approach is that we can calculate the free energy of the model, which is proportional to the language model probability, very fast. Therefore, we are able to use the RBM-based language model during decoding and not only in the rescoring phase. The model is described in detail in [16].

The RBM used for the language model consists of two layers, which are fully connected. In the input layer, for every word position there are as many nodes as words in the vocabulary. Since we used an 8-gram language model, there are 8 word positions in the input layer. These nodes are connected to the 32 hidden units in the hidden layer.

During decoding, we calculate the free energy of the RBM for a given n-gram. The product of this values is then used as an additional feature in the log-linear model of the decoder.

9. Postprocessing for Agreement Correction

The agreement in gender and number is one of the challenging problems encountered when translating from English into a morphologically richer language such as French. Consequently, a special postprocessing was designed in order to remedy the case where disagreements between nouns and related surrounding words exist. This post-processing is based on the POS tags generated by LIA tagger[3]. In order to improve the agreement features, several post-processing heuristics are applied on a sentence basis, which include the correction of the grammatical number and gender of adjective, article, possessive determiner, forms of *quelque* and past participles based on their corresponding nouns.

In order to minimize spurious assignments when finding instances of these parts of speech related to a specific noun, strict heuristics are used: Adjectives must appear straight before or after the noun. Articles, possessive determiners and forms of *quelque* have to directly precede nouns or have at most one adjective in between. Past participles must stand after (possibly reflexive) inflected forms of *être* that immediately follow nouns.

10. Results

In this section, we present a summary of our experiments for all tasks we have carried out for the IWSLT 2012 evaluation. It includes the official systems for the MT and SLT translation tasks and additional systems for other language pairs: English-German, English-Chinese and English-Arabic translations. All the reported scores are the case-sensitive BLEU, and calculated based on the provided Dev and Test sets.

[3] http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/bechet/download_fred.html

10.1. MT Task

Table 1 summarises how our MT system evolved. The baseline translation model was trained on EPPS, TED, NC, and Giga corpora. This big model was adapted with a smaller one trained on TED data only as described in Section 5. The language model is a log-linear combination of three language models trained on different data sets: the French side of the EPPS, TED, and NC corpora, the provided monolingual news data (Monolingual EPPS, NC and News Shuffled), and a smaller in-domain language model trained on TED data. The reordering in this system was handeled as a preprocessing step using POS-based rules as described in Section 4. The result of this setting was 28.5 BLEU points on Dev and 31.73 on Test. The performance could be improved by around 0.4 on Dev and 0.2 on Test by using a bilingual language model (details about bilingual language model computation can be found in [17]). An additional 0.2 on both Dev and test could be gained by using a cluster language model where the clusters were trained on the in-domain TED data. After that, changing the adaptation strategy by the union selection discussed in Section 5 shows slight improvement of 0.1 on both Dev and Test. The effect of the DWL trained on only the TED corpus was rather dissimilar on Dev and Test. While it slightly improved the score on Dev (0.1) it has a much greater effect on Test (0.5). Further small improvement could be observed by using a continuous space language model: around 0.09 on both Dev and Test. Finally, by using the POS-based post-processing correction of the agreement on the target side the score on Test could be improved by an additional 0.06, resulting in 32.84 BLEU points on Test. We submitted the translations of Test2011 and Test2012 generated by this final system as primary; the translations generated by the second best system (same as the final but without agreement corrections) as contrastive.

System	Dev	Test
Baseline	28.50	31.73
+Bilingual LM	28.93	31.90
+Cluster LM	29.15	32.13
+CSUnion	29.27	32.21
+DWL	29.37	32.70
+RBM LM	29.46	32.78
+Agreement Correction	-	32.84

Table 1: Summary of experiments for the English-French MT task

10.2. SLT Task

The baseline system of the speech translation task used almost the same configuration as the one for the MT task, for which the POS-based reordering and the adaptation for both translation and language model with TED data were added to the baseline. The special processing we have done for SLT

task lie in the following aspects.

In order to simplify building the system, we did not retrain a new alignment for the SLT task, but modify the phrase tables from the MT task to make it suitable for the SLT task. Casing information and punctuation except periods has been removed from the source side of the phrase table. Then we feed this new phrase table with possibly duplicate phrase pairs into the SLT system and let the decoder select the best ones for a translation. For the purpose of comparison, we also rebuild a whole new SLT system, in which the alignment, the phrase table and all other models are newly generated with the training data without punctuation and casing information. However, the newly trained system is not better than the MT system with the modified phrase table. The experimental results are presented in Table 2. **large-retrain-PT** are with the newly trained phrase table on the same corpora. **large-modify-PT** is the system with the modified phrase table trained on bilingual corpora TED, NC, EPPS and Giga corpus. We can see that the completely retraining the system does not improve the result. It is very surprising that the retrained system hurts the result much. One possible explanation could be punctuations are very help to generate good alignments. In order to know the reasons more clearly, more experiments should be done in the future.

Another difference to the MT system is the the data used to build translation model does not include the Giga corpus. It includes only TED, NC and EPPS, since including the Giga corpus could not improve the translation results in the SLT task, as it does in the MT task. The intermediate experiments of comparing these two training data sets are shown in Table 2. **small-modify-PT** is the system trained only on TED, NC and EPPS. The systems trained on TED, NC, EPPS and Giga are called **large**.

System	Dev	Test(ASR)
large-retrain-PT	17.14	18.92
large-modify-PT	18.67	21.08
small-modify-PT	18.93	21.84

Table 2: Intermediate experiments with different phrase tables for the English-French SLT task

Our SLT system is optimized on the modified Dev text data by removing the punctuation except periods and lowercasing. And we have tested the system both on modified text test data which is with the same processing as the Dev text data and on the ASR output of the test data. Table 3 presents the results optimized on modified Text and ASR output, respectively. The two columns marked with **Test(ASR)** are comparable scores. There is no convinced evidence that on which condition the optimization is better. In the settings of "Baseline", "Adaptation" and "Bilingual LM" optimizing on ASR output gets better results. After applying all models, the system optimized on the modified text data wins about 0.5 BLEU points. Considering the final result after adding all models is better and the test data from modified Text if more reliable than the ASR output, we have chosen the system optimized on the modified text data as our primary system.

We present our system for the SLT task step by step in Table 3. The bilingual language model was trained on the EPPS corpus and all other available parallel data, whose punctuation marks on the source side are all removed. The cluster language model is trained on the TED corpus, where the words are classified into 50 classes. The DWL model is also trained on the TED corpus, but the punctuation and casing information have been removed from the source side of the training data.

Compared to the baseline the SLT system has improved about 1.1 BLEU on both text and ASR test data by adding all the models. The largest gain is about 0.5 by adding the cluster-based language model. The domain adaptation model has improved all scores on Dev, text Test and ASR Test. It especially improves the text Test by 0.5 BLEU. The bilingual language model does not seem to contribute much to the results, except a little improvement of 0.2 on the ASR test data. Then we add the DWL model which also improves the test data by about 0.2 BLEU points. Finally we have carried out the morphology agreement correction as described in Section 9, which improves around 0.1 on the test data.

This system was the system we used to translate the SLT evaluation set for our submission. We have submitted one primary system and three contrastive systems. The primary system is the translation of the ASR output *system1* with all models presented in Table 3. And the contrastive systems are the translations of the ASR outputs *system1* - *system3* excluding the **Agreement Correction** model.

10.3. Additional Language Pairs

10.3.1. English-German

Several experiments were conducted for the English-German MT track on the TED corpus. They are summarized in Table 4. The baseline system is essentially a phrase-based translation system with some preprocessing steps on both source and target sides. Adapting huge parallel data from EPPS and NC to TED translation model helps us gain 0.71 BLEU scores on the test set. Short-range reordering based on POS information yields reasonable improvements on both development and test sets by about 0.5 BLEU points. In the language modeling aspect, different factors were experimented with, and 4-gram POS language model using RF-Tagger[4] slightly improves our system over the development set by 0.22 BLEU points but considerably shows its impact on test set with an improvement of 1 BLEU point. We approach our best system by adding a 9-gram cluster-based language model where the German side corpus is grouped into 50 classes, yielding 22.61 and 22.93 BLEU points on development and test sets, respectively.

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger/

	Optimization on Text			Optimization on ASR	
System	Dev (Text)	Test (Text)	Test (ASR)	Dev (ASR)	Test (ASR)
Baseline	25.37	27.57	21.68	19.11	21.86
+ Adaptation	25.64	28.08	21.90	19.31	22.04
+ Bilingual LM	25.07	28.08	22.07	19.14	22.28
+ Cluster LM	25.17	28.79	22.57	19.32	22.40
+ DWL	25.06	28.84	22.79	19.34	22.23
+ Agreement Correction	-	-	22.86	-	-

Table 3: Summary of experiments for the En-Fr SLT task

System	Dev	Test
Baseline	20.59	20.50
+ Adaptation	21.39	21.21
+ Reordering	21.97	21.74
+ POS LM	22.19	22.73
+ Cluster LM	22.61	22.93

Table 4: Experiments for the English-German on TED task

In this English-German translation system, we have also tried some other models such as using DWL, long-range reordering, bilingual language model as well as external monolingual language models but we do not gain noticeable improvements. Moreover, some experiments on tree-based reordering, which we believe helpful in this language pair, has been reserved for further considerations due to the limited time.

10.3.2. English-Chinese

With the bilingual data released by the TED Task of IWSLT 2012 we have developed an English-Chinese translation system. As it is an initial system for this new translation direction, we have made the main effort on data processing and preprocessing.

There are three corpora that could be used: the TED bilingual sentence-aligned corpus, the UN bilingual document-aligned corpus and the monolingual Google Ngrams corpus. In our system we have used the TED corpus to train the translation model and trained a language model on TED, UN and Google Ngrams. In addition we classify the Google Ngram corpus with its year information, such as google1980 is the ngrams from 1980-1989, and train a language model separately on each class. Our experience has shown that google1980 has contributed the most to the improvement, even more than the whole Google Ngram corpus.

In constrast to European languages, there are no spaces between Chinese words. Therefore, in the preprocessing of English-Chinese translation we need to decide on whether to segment Chinese into words, or to segment it into characters. We have tried both in our experiments. For the Chinese word segmentation we have made use of the Stanford Chinese word segmenter[5]. For the Chinese character segmentation we have simply inserted a space between neighbor Chinese characters. Then we have trained two systems: one based on Chinese words, the other based on Chinese characters. Table 5 shows the results from the two systems. Since the evaluation scores on Chinese words (**Test(Word)**) and on Chinese character (**Test(Cha.)**) are not comparable to each other, we segment the translation hypothesis on words into Chinese characters. Then the scores at the two columns **Test(Cha.)** are comparable. We can see that the system trained on characters is usually better than the system on words.

In Table 5 we present the steps which achieve improvement. The baseline system is trained only on the TED corpus (both for translation model and language model). By adding all possible language models and a reordering model, the BLEU score on test data has gained 0.2 points in total. Most improvements come from the larger language model. It seems that the current reordering model does not work quite well for the English-Chinese translation. Further analysis and work need to be done on the reordering model.

System	on characters		on words	
	Dev (Cha.)	Test (Cha.)	Test (Cha.)	Test (Word)
Baseline(4gram LM)	14.37	17.26	16.69	9.92
8gram LM	14.48	17.28	17.08	10.03
+ 4gram UN LM	14.61	17.38	16.80	9.99
+ POS Reordering	14.69	17.28	17.32	10.23
+ 5gram google1980	14.73	17.47	16.82	9.84

Table 5: Translation results for English-Chinese

The other models that we have tried, but have not given improvement to the system, include sentence-aligned extraction from the UN corpus and long-range reordering as described in [18].

[5] http://nlp.stanford.edu/software/segmenter.shtml

10.3.3. English-Arabic

The parallel data provided for this direction was from TED and UN. As for the English-Chinese direction (presented in Section 10.3.2), greater effort was devoted to the data preprocessing. The preprocessing for the English side is identical to the one used in the English-French system of the MT Task. Some of these preprocessing operations, such as long pair removal, were also applied to the Arabic side. In addition to that, the Arabic side was further orthographically transliterated using Buckwalter transliteration [19]. Tokenization and POS tagging were performed by the AMIRA toolkit [20]. The resulting translation is converted back to Arabic scripting before evaluation.

Table 6 presents some initial experiments for the English-Arabic pair. The baseline system uses only TED data for translation and language modeling. This gave a score of 13.12 on Dev and 8.05 on Test. This system was remarkably enhanced by introducing the short range reordering rules. The scores were improved by about 0.3 on Dev and 0.2 on Test. Adding monolingual data from the UN corpus had a great impact on the score on Dev (improved by 0.6), whereas it has a much lower effect on Test (improves by 0.1 only). In this last setting, three language models were log-linearly combined: one trained on TED data, one trained on UN data, and another one trained on both. Since the UN corpus was provided as raw data (no sentence alignment was performed before), we selected a sub-corpus of documents consisting of exactly the same number of sentences. This resulted in around 500K additional parallel sentences. The line **SubUN parallel** in Table 6 shows that these data had almost no effect on the system's performance. It increased the score on Dev by 0.02 and by 0.07 on Test. However, using the first translation model (trained on TED only) as indomain model to adapt the last setting shows slightly better improvements (around 0.1 on Dev and Test). Using a bilingual language model rather harmed the system on Dev by around -0.1 but improved the score on Test by 0.06. We choose to include this model because combined with the cluster language model it could improve our system by around 0.2 on Dev and Test wheras none of these models alone could outperform this score (some of these experiments are not reported here).

System	Dev	Test
Baseline	13.12	8.05
+ POS Reordering	13.46	8.23
+ Language models	14.08	8.32
+ SubUN parallel	14.10	8.39
+ TM Adapt	14.24	8.46
+ Bilingual LM	14.15	8.52
+ Cluster LM	14.28	8.63

Table 6: Experiments for the English-Arabic

11. Conclusions

In this paper, we presented the systems with which we participated in the TED tasks in both speech translation and text translation from English into French in the IWSLT 2012 Evaluation campaign. Our phrase-based machine translation system was extended with different models.

For the official language pair, even though we were authorized to use the UN parallel corpus and the monolingual Google Books Ngrams, these data had always a negative impact on our system's quality. More experiments should be carried out to extract some useful parts of these large data.

The successful application of different supplementary models trained exclusively on TED data (cluster language model, DWL, and continuous space language model) shows the usefulness and importance of in-domain data for such tasks, regardless of their small size.

The large amount of data used to train the different models integrated in our statistical system could not compensate for the ambiguity of translating into a morphologically richer language. Therefore, applying very simple and limited heuristics based on the target language grammar gave small but consistent improvments using the POS-based agreement correction.

We also presented experiments with several additional pairs. Namely, from English into one of the languages German, Chinese, or Arabic.

The use of additional bilingual corpora on adapting translation models as well as more complicated features from different language models led to expected performance in the English-German translation system. The effects of other techniques, e.g. long-range reordering or discriminative word alignment (DWA), were less obvious, mainly coming from the characteristics of the TED data.

In case of English-Chinese, we have found that the system based on Chinese characters works better than the system based on Chinese words. The BLEU score calculated on Chinese characters and Chinese words are also different: the BLEU score on character is about 17 while evaluation on the words the score is around 10. In addition we found that the current reordering model does not help much on this language pair. Further work needs to be done in this field in the future.

Due to the limited amount of data, the English-Arabic system performed relatively poorly. Furthermore, it showed eventual discrepency between Dev and Test data. Here again, as mentioned before, the UN data were not helpful.

12. Acknowledgements

This work was partly achieved as part of the Quaero Programme, funded by OSEO, French State agency for innovation. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287658.

13. References

[1] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 Evaluation Campaign," in *Proc. of the International Workshop on Spoken Language Translation*, Hong Kong, HK, December 2012.

[2] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[3] M. Mediani, E. Cho, J. Niehues, T. Herrmann, and A. Waibel, "The KIT English-French Translation systems for IWSLT 2011," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, 2011.

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

[5] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[6] 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.

[7] M. Mediani, J. Niehues, and A. Waibel, "Parallel Phrase Scoring for Extra-large Corpora," in *The Prague Bulletin of Mathematical Linguistics*, no. 98, 2012, pp. 87–98.

[8] G. F. Foster, R. Kuhn, and H. Johnson, "Phrasetable smoothing for statistical machine translation," in *EMNLP*, 2006, pp. 53–61.

[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. 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.

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

[12] 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.

[13] F. J. Och, "An Efficient Method for Determining Bilingual Word Classes." in *EACL'99*, 1999.

[14] A. Mauser, S. Hasan, and H. Ney, "Extending Statistical Machine Translation with Discriminative and Trigger-based Lexicon Models," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, ser. EMNLP '09, Singapore, 2009.

[15] H.-S. Le, A. Allauzen, and F. Yvon, "Continuous Space Translation Models with Neural Networks," in *Proceedings of the 2012 Conference of the NAACL-HLT*, Montréal, Canada, June 2012.

[16] J. Niehues and A. Waibel, "Continuous Space Language Models using Restricted Boltzmann Machines," in *Proceedings of the nineth International Workshop on Spoken Language Translation (IWSLT)*, 2012.

[17] J. Niehues, T. Herrmann, S. Vogel, and A. Waibel, "Wider Context by Using Bilingual Language Models in Machine Translation," in *Sixth Workshop on Statistical Machine Translation (WMT 2011)*, Edinburgh, UK, 2011.

[18] J. Niehues and M. Kolss, "A POS-Based Model for Long-Range Reorderings in SMT," in *Fourth Workshop on Statistical Machine Translation (WMT 2009)*, Athens, Greece, 2009.

[19] N. Habash and F. Sadat, "Arabic Preprocessing Schemes for Statistical Machine Translation," in *Proceedings of the NAACL-HLT*, ser. NAACL-Short '06, Stroudsburg, PA, USA, 2006.

[20] M. Diab, "Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking," in *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April 2009.