The KIT Translation Systems for IWSLT 2013

Thanh-Le Ha, Teresa Herrmann, Jan Niehues, Mohammed Mediani,
Eunah Cho, Yuqi Zhang, Isabel Slawik and Alex Waibel

Institute for Anthropomatics
KIT - Karlsruhe Institute of Technology
firstname.lastname@kit.edu

Abstract

In this paper, we present the KIT systems participating in all three official directions, namely English→German, German→English, and English→French, in translation tasks of the IWSLT 2013 machine translation evaluation. Additionally, we present the results for our submissions to the optional directions English→Chinese and English→Arabic.

We used phrase-based translation systems to generate the translations. This year, we focused on adapting the systems towards ASR input. Furthermore, we investigated different reordering models as well as an extended discriminative word lexicon. Finally, we added a data selection approach for domain adaptation.

1. Introduction

In the IWSLT 2013 Evaluation Campaign [1], we participated in the tasks for text and speech translation for all the official language pairs: English→German, German→English and English→French as well as two optional directions. The TED tasks consist of automatic translation of both the manual transcripts (MT task) and transcripts generated by automatic speech recognizers (SLT task) for talks held at the TED conferences[1]. For German→English, the test data was collected from the TEDx project[2].

The TED talks are given in English in a large number of different domains. Some of these talks are manually transcribed and translated by global volunteers into many languages [2]. The TED translation tasks this year bring up interesting challenges: (1) the problem of adapting general models - mainly trained on news data - towards the diverse topics in TED talks, (2) the need of universal techniques for translating texts from and to various languages, and (3) the appropriate solution for inserting punctuation marks and case information on automatic speech recognition (ASR) outputs for the spoken language translation (SLT) task.

To deal with those challenges, we provided several advanced adaptation methods both for translation and language models to leverage both the wide coverage of large data portions and the domain-relevance of the TED corpus. In addition, we optimized our universal techniques to better conform with different languages.

Compared to our last year's system, we focused on four new components: handling of ASR input (Section 3), combination of reordering models of different linguistic abstraction levels (Section 4), data selection for language model (LM) adaptation (Section 5) and an extended discriminative word lexicon (Section 6).

The next section briefly describes our baseline system, while Sections 3 through 7 present the different components and extensions used by our phrase-based translation system. After that, the results of the different experiments, including official and optional language pair systems, are presented and we close the paper with a conclusion.

2. Baseline System

Among the parallel data provided, we utilize EPPS, NC, TED, Common Crawl for English→German and German→English, plus Giga for English→French. The monolingual data we used include the monolingual part of those parallel data, the News Shuffle corpus for all three directions and additionally the Gigaword corpus for English→French and German→English.

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. In German→English, we also apply compound splitting [3] to the source side of the corpus. Furthermore, an SVM classifier is used to filter out the noisy sentence pairs in the Giga English→French corpus and the Common Crawl as described in [4].

Unless stated otherwise, the language models used are 4-gram language models with modified Kneser-Ney smoothing, trained with the SRILM toolkit [5] and scored in the decoding process with KenLM [6]. The word alignment of the parallel corpora is generated using the GIZA++ Toolkit [7] for both directions. Afterwards, the alignments are combined using the grow-diag-final-and heuristic. For German→English, we use a discriminative word alignment (DWA) approach [8]. The phrases are extracted using the

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

Moses toolkit [9] and then scored by our in-house parallel phrase scorer [10]. Phrase pair probabilities are computed using modified Kneser-Ney smoothing as in [11].

In all directions, beside the word-based language models, some of the non-word language models are used. In order to increase the bilingual context used during the translation process, we use a bilingual language model as described in [12]. To model the dependencies between source and target words even beyond borders of phrase pairs, we create a bilingual token out of every target word and all its aligned source words. The tokens are ordered like the target words. In addition, to alleviate the sparsity problem for surface words, we use a cluster language model based on word classes. This is done in the following way: In a first step, we cluster the words of the corpus using the MKCLS algorithm [13]. Then we replace the words in the TED corpus by their cluster IDs and train an n-gram language model on this corpus consisting of word classes.

3. Preprocessing for Speech Translation

The system translating automatic transcripts needs special preprocessing on the data, since generally there is no or no reliable case information and punctuation in the automatically generated transcripts. We have used a monolingual translation system as shown in [14] to deal with the difference in casing and punctuation between a machine translation (MT) and an SLT system. In contrast to the condition in their work, in this evaluation campaign sentence boundaries are present in the test sets. Therefore, we use this monolingual translation system for predicting commas instead of all punctuation marks in the test set. In addition to predicting commas, we also predict casing of words using the monolingual translation system. This preprocessing will be denoted as Monolingual Comma and Case Insertion (MCCI).

In order to build the monolingual system which translates a source language into the same language with commas inserted, we prepare the parallel corpus for training. For the source side of the corpus, we take the preprocessed monolingual corpus of a normal translation system, remove all punctuation marks, and insert a period mark at the end of each line. For the target side of the corpus, we take the preprocessed corpus of same language from the normal translation system and replace all sentence-final punctuation marks such as "!", "?", "." by a period. Therefore, the only difference between the source and the target side corpus is inserted commas on the target side.

In this evaluation campaign we work with two source languages, English and German. Therefore, we build a monolingual translation systems each for the two languages. The speech translation system with English on the source side is built using true-cased English source and target side. As the test set often contains only lower-cased letters, in the English monolingual system we take this already lower-cased, preprocessed automatic transcript for translation. In order to match this input during decoding, the source side of a phrase table is lower-cased. As the case information contains more information for German, the German monolingual translation system is built using lower-cased German source and true-cased target side. All words in the preprocessed German automatic transcript are lowercased, but are translated into true-cased text using the monolingual translation system.

The monolingual translation systems for both languages are built on the corresponding side of the EPPS, TED, and NC corpus, which sum up to 2.2 million sentences. A 4-gram language model trained on the word tokens is used. Word reordering is ignored in these systems. In order to capture more context, we use a 9-gram language model trained on part-of-speech (POS) tokens. Moreover, a 9-gram cluster language model is trained on 1,000 clusters, based on the MKCLS algorithm as described in the baseline system.

For the speech translation tasks, the output of the monolingual translation system becomes the input to our regular translation system which is trained using data with punctuation marks.

4. Word Reordering Model

Word reordering is modeled in two ways. The first is a lexicalized reordering model [15] which stores reordering probabilities for each phrase pair. The second model consists of automatically learned rules based on POS sequences and syntactic parse tree constituents and performs source sentence reordering according to target language word order.

The rules are learned from a parallel corpus with POS tags [16] for the source side and a word alignment to learn continuous reordering rules that cover short-range reorderings [17]. Discontinuous rules consist of POS sequences with placeholders and allow long-range reorderings [18]. In addition, we apply a tree-based reordering model [19] to better address the differences in word order between German and English. Syntactic parse trees [20, 21] for the source side of the training corpus and a word alignment are required to learn rules on how to reorder the constituents in the source sentence to simulate target sentence word order. The POS-based and tree-based reordering rules are applied to each input sentence before translation. The resulting reordered sentence variants as well as the original sentence are encoded in a word lattice.

In order to apply the lexicalized reordering model, the lattice includes the original position of each word. Then the lattice is used as input to the decoder. During decoding the lexicalized reordering model provides the reordering probability for each phrase pair. At the phrase boundaries, the reordering orientation with respect to the original position of the words is checked. The probability for the respective orientation is included as an additional score in the log-linear model of the translation system.

5. Adaptation

In order to achieve the best performance on the target domain, we perform adaptation for translation models as well as language models.

We adapt the translation model (TM) by using the scores from the in-domain and out-of-domain phrase table as described in the backoff approach [22]. This results in a phrase table with six scores, the four scores from the general phrase table as well as the two conditional probabilities from the in-domain phrase table. In addition, we adapt the candidate selection in some of our systems by taking the union of the candidates translations from both phrase tables (CSUnion).

The language model (LM) is adapted by log-linearly combining the general language model and an in-domain language model trained only on the TED data. In addition, in some of the systems we combine these language models with a third language model. This language model was trained on data automatically selected using cross-entropy differences [23]. We selected the top 5M sentences to train the language model.

6. Discriminative Word Lexica

Mauser et al. [24] have shown that the use of DWL can improve the translation quality. For every target word, they train a maximum entropy model to determine whether this target word should be in the translated sentence or not using one feature per source word. In our system we use the extended version using also source context and target context features [25]. When using source context features, not only the words of the sentence are used as features, but also the n-grams occurring in the sentence. The target context features encode information about the surrounding target words.

One specialty of the TED translation task is that we have a lot of parallel data we can train our models on. However, only a quite small portion of these data, the TED corpus, is very important for the translation quality. Therefore, we achieve a better translation performance by training the models only on the TED data.

7. Continuous Space Language Model

In recent years, different approaches to integrate continuous space models have shown significant improvements in the translation quality of machine translation systems [26]. Since the long training time is the main disadvantage of this model, we only train 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 the free energy of the model, which is proportional to the language model probability, can be calculated very efficiently. Therefore, we are able to use the RBM-based language model during decoding and not only in the rescoring phase.

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 a 4-gram language model, there are 4 word positions in the input layer. These nodes are connected to 32 hidden units in the hidden layer. The model is described in detail in [27].

8. Results

In this section, we present a summary of our experiments for all tasks we have carried out for the IWSLT 2013 evaluation. All the reported scores are case-sensitive BLEU scores calculated based on the provided development and test sets.

8.1. English→German

We conducted several experiments for English→German translation using the available data. They are summarized in Table 1. The baseline system is a phrase-based translation system using POS-based reordering rules. Preprocessing of the source and target language of the training corpora is performed as described above. Adaptation of the phrase table and language model using the in-domain part of the training data is included, as well as a bilingual language model to increase the source context across phrase boundaries. Finally, the baseline system also includes a cluster-based language model using the clusters automatically generated by the MKCLS toolkit.

System	Dev	Test
Baseline	23.58	23.50
+ Tree-based Rules	23.61	23.87
+ Lexicalized Reordering	23.74	23.93
+ POSLM	23.81	24.14
+ DWL	24.44	24.76
+ Class-based 9-gram LMs	24.19	24.93
+ TargetContext + LM DataSelection	24.24	**25.06**

Table 1: Experiments for English→German (MT)

By adding tree-based reordering rules and a lexicalized reordering model we increase the translation quality by more than 0.4 BLEU points. An additional language model for POS sequences gives another increase of 0.2 BLEU points. A remarkable improvement of 0.6 can be observed by introducing a discriminative word lexicon trained on the in-domain data where bigrams are used to include more information about the context words on the source side. Extending the class-based language model to 9-grams leads to further improvement by 0.2. The final system includes target context features in the discriminative word lexicon and a language model trained on 5 million sentences selected from all data based on cross entropy similarity.

8.1.1. SLT Task

For the English→German SLT task, we used one of the systems developed for the MT task. For reordering, it includes the lexicalized reordering model and long-range reordering rules. The tree-based rules are excluded since they do not conform well with the speech data. In addition, the system uses 9-gram POS-based and MKCLS language models and an in-domain DWL with source context. This system ignores case information on the source side. While both development and test data were available for the MT task, for the SLT task only one data set was provided. Therefore, we used it for testing and performed optimization on text data.

In order to adapt the system further towards the task of translating speech input, we added the monolingual comma and case insertion model, which performs a preprocessing step consisting of monolingual translation of lowercased English speech into true-cased English while also inserting commas. For this, no new optimization was performed, only the input was changed. This special treatment of the speech input helped improve the system performance by 1.3 BLEU points. Table 2 shows the overview of the speech translation system.

ASR Adaptation	Test
Baseline	17.60
MCCI	**18.92**

Table 2: Experiments for English→German (SLT)

8.2. German-English

We summarize the development of the German→English system in Table 3. The translation model of the baseline system uses a bilingual language model. It uses all types of reordering rules and a lexicalized reordering model. Furthermore, three language models are combined log-linearly in this system. One language model is trained on all data, one only on the in-domain data and we use one cluster language model trained on all data using 1,000 clusters. Adding the DWL trained on the TED corpus using source and target context features improves the performance by 0.9 BLEU points. Further improvements are achieved by adding a language model trained on the automatically selected data. We further adapt the system to the TED task using the union candidate selection and by adding a RBM-based language model. This improves the system only slightly by 0.1 BLEU points. Finally, we replace the cluster language model by one trained only on the TED corpus and also use morphological operations to translate unknown word forms [12].

8.2.1. SLT Task

For the SLT task, we use the MT system without the in-domain cluster LM and morphological operations. By directly using the MT system to translate the ASR output, a translation quality of 18.33 BLEU points is reached. As there are often no case information and commas in the ASR output, we remove these information from the source side of the phrase table. Using this system, we improve the translation quality to 19.09. Then we use the MCCI system described in Section 3 to insert case information and commas into the ASR output. When translating this modified ASR output, we reach a final BLEU score of 20.1.

System	Dev	Test
Baseline	35.17	29.76
+ DWL	35.42	30.65
+ LM DataSelection	35.51	30.80
+ CSUnion + RBMLM	35.75	30.87
+ In-domain Cluster LM	35.74	31.10
+ Morphological Operations	-	**31.15**

Table 3: Experiments for German→English (MT)

ASR Adaptation	Test
Baseline	18.33
Phrase Table	19.09
MCCI	**20.10**

Table 4: Experiments for German→English (SLT)

8.3. English→French

Table 5 reports some remarkable improvements as we combined several techniques on the English→French direction. The big phrase table is trained on TED, EPPS, NC, Giga and Crawl data, while the language model is trained on the French part of those corpora plus News Shuffle. The system also uses short-range reordering rules derived from smaller data portions (TED, EPPS and NC). The result of this setting is 31.08 BLEU points.

System	Dev	Test
Baseline	27.68	31.08
+ PT+LM Adaptation	28.48	31.76
+ Bilingual LM	28.66	32.57
+ POS+Cluster LMs	28.85	32.53
+ Lexicalized Reordering	29.22	32.83
+ DWL Source Context	29.45	**33.06**

Table 5: Experiments for English→French (MT)

Several advanced adaptations are conducted both on translation and language models. First, the phrase table is adapted using the clean EPPS, NC and TED data. Afterwards, it is adapted towards the TED domain. For the language models, we follow the similar adaptation scheme with the models ranging from in-domain to general-genre data.

We log-linearly combine the language models trained on TED, EPPS, NC, Giga, and Crawl by minimizing the perplexity on the development set. Those adaptation techniques boost the system around 0.7 BLEU points. Further gains come from using different non-word language models. Introducing the bilingual language model leads to a small improvement of 0.18 on Dev and 0.81 BLEU points on Test. Adding a 9-gram POS-based language model and a 4-gram 50-cluster language model trained on in-domain data helps gain almost 0.2 BLEU points on Dev, but results in a slightly reduction of 0.04 on Test. The system is further enhanced by 0.3 BLEU points when we integrated lexicalized reordering probabilities as an independent feature. Finally, by taking the source context of the DWL into account, we achieve the best system with a 0.23 increase, reaching 33.06 BLEU points.

8.3.1. SLT Task

We approached the SLT tasks in two distinct ways. The first is that we use the best system of the MT task to translate the ASR outputs which were already preprocessed by Monolingual Comma and Case Insertion (MCCI) system as mentioned in Section 3. The second approach is the system named ASR-Dedicated, which evolves from rebuilding the translation model from modified Giza alignments dedicated for ASR data only. The modifications consist of removing the case and punctuation marks except the period.

Table 6 presents the results using the best MT system to translate two ASR outputs and from the second approach. The ASR outputs are the raw text without any comma (None) and the output using MCCI preprocessing. The numbers show that a big improvement of almost 3 BLEU points comes from the input preprocessed by MCCI. The commas MCCI inserted have a great effect on the fluency of the ASR output and consequently improved the translation quality. The numbers also show that the system trained and optimized to work best for texts would work adequately for ASR outputs as well.

We submitted the best MT system with MCCI as the primary, and the second approach's result as the contrastive.

ASR Adaptation	Test
None	20.75
MCCI	**23.69**
ASR-Dedicated	22.90

Table 6: Experiments for English→French (SLT)

8.4. English→Arabic

For this pair, we use the parallel data from TED. The UN parallel data is provided in raw format. In order to get useful parallel pairs out of this raw data, we segment the two sides into sentences, exclude all documents having a large difference in number of sentences, sentence-align the resulting document pairs, and finally filter out the noisy sentence pairs.

We use the default sentence segmenter provided by the NLTK toolkit [28] to segment both sides. The sentence alignment is performed using the Hunalign aligner [29]. Since this aligner works better with a lexicon, we build one from Giza alignments trained on the TED corpus. The filtering is carried out using an SVM classifier as stated in Section 2. The tokenization and POS tagging of the Arabic side are performed using the AMIRA toolkit [30].

In addition to the parallel data provided, the fifth edition of the LDC Gigaword Arabic corpus is also used for language modeling.

Table 7 summarizes the experiments for the English→Arabic pair. The baseline translation model is trained on all parallel data (TED and UN) and involves many language models which are log-linearly combined. These include individual models one from each corpus (TED, UN, Gigaword) and two more (UN & TED and all corpora together). In this configuration we use the short range reordering. This system gives 13.15 on Dev and 8.43 on Test. The effect of translation model adaptation is remarkable: it improves the system performance by almost 1.4 BLEU on Dev and 0.26 on Test. Slight improvements could be brought by introducing more language models. For instance, using a bilingual language model trained on all parallel data increases the performance on Dev by almost 0.2 while it has no observable effect on Test. On the other hand, adding a 4-gram cluster language model trained on TED only (with 50 classes) enhances the score on Test by 0.2 while it leaves the Dev score almost unchanged. This last system is used in our submission.

System	Dev	Test
Baseline	13.15	8.43
+ PT Adaptation	14.54	8.69
+ Bilingual LM	14.79	8.70
+ Cluster LM	14.81	**8.92**

Table 7: Experiments for English→Arabic (MT)

8.5. English→Chinese

The English→Chinese system is trained on the bilingual TED and filtered UN corpora. As the UN corpus is document-aligned, we have filtered out about 30k aligned sentences as training data with a KM algorithm. The weight of a sentence pair is the accumulation of word and its translation occurring in a dictionary. The dictionary used here is from LDC (LDC2002L27). The language models are trained on the monolingual TED data and the target side of the whole UN data.

In contrast to European languages, there are no spaces between Chinese words. In our primary system we segment Chinese into characters and tokenize and lowercase English.

Adaptation, reordering and DWL source context models have given contribution to the improvement of translation. In Table 8 we present the steps which achieve improvement. The baseline is a monotone translation with 6-gram language model. As the adaptation described in Section 5, we use the TED corpus as the in-domain data to adapt the phrase table and language model. We use two reordering models: short-range POS-based reordering and lexicalized reordering, which are described in Section 4. Finally, after adding the DWL source context model as described in Section 6 and CSUnion model in Section 5, the BLEU score on test data has gained more than 1 point compared to the baseline.

We have also built a system based on Chinese words as a contrastive system, where the words are generated with the Stanford word segmenter[3].

System	MT		SLT
	Dev	Test	Test
Baseline	14.01	16.75	-
+ Adaptation	14.61	16.77	-
+ POS Reordering	14.71	17.51	-
+ Lexicalized Reordering	14.91	17.18	-
+ DWL+CSUnion	15.14	**17.84**	**17.28**

Table 8: Experiments for English→Chinese

8.5.1. SLT Task

The speech translation system has used the same configuration as the best one for the MT task. We built the test data set by removing the case information and punctuation from the text test data. In order to apply the system trained on text for speech automatic transcripts, we predict commas with the preprocessing described in Section 3. The result is shown in Table 8.

9. Conclusions

In this paper, we presented the systems with which we participated in the TED tasks in both speech translation and text translation of the IWSLT 2013 Evaluation Campaign. Our phrase-based machine translation system was extended with different models.

When translating ASR input, we need to adapt the system to these conditions. Often case information or commas are missing or misplaced. Therefore, we use a method to automatically correct this information in order to directly use our default translation model without training a separate model.

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. Furthermore, we could adapt the system even more to the task by using data selection methods.

The DWL allows us to include arbitrary features when calculating the translation probabilities. By extending these models to also include contextual information about the source and target sentence, we were able to increase the translation performance. Furthermore, we could improve the translation performance by combining information about the word order from different linguistic levels.

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

10. Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287658.

11. References

[1] M. Cettolo, J. Niehues, S. Stueker, L. Bentivogli, and M. Federico, "Report on the 10th IWSLT Evaluation Campaign," in *IWSLT 2013*, 2013.

[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] P. Koehn and K. Knight, "Empirical Methods for Compound Splitting," in *EACL*, Budapest, Hungary, 2003.

[4] 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.

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

[6] K. Heafield, "KenLM: faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, United Kingdom, July 2011, pp. 187–197.

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

[8] J. Niehues and S. Vogel, "Discriminative Word Alignment via Alignment Matrix Modeling." in *Proceedings of Third ACL Workshop on Statistical Machine Translation*, Columbus, USA, 2008.

[9] 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.

[10] 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.

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

[12] 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.

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

[14] E. Cho, J. Niehues, and A. Waibel, "Segmentation and Punctuation Prediction in Speech Language Translation using a Monolingual Translation System," in *Proceedings of the Ninth International Workshop on Spoken Language Translation (IWSLT)*, 2012.

[15] P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot, "Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Pittsburgh, PA, USA, 2005.

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

[17] 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.

[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] T. Herrmann, J. Niehues, and A. Waibel, "Combining Word Reordering Methods on different Linguistic Abstraction Levels for Statistical Machine Translation," in *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*. Altanta, Georgia, USA: Association for Computational Linguistics, June 2013.

[20] A. N. Rafferty and C. D. Manning, "Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines," in *Proceedings of the Workshop on Parsing German*, 2008.

[21] D. Klein and C. D. Manning, "Accurate Unlexicalized Parsing," in *Proceedings of ACL 2003*, 2003.

[22] J. Niehues and A. Waibel, "Detailed Analysis of Different Strategies for Phrase Table Adaptation in SMT," in *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*, 2012.

[23] R. Moore and W. Lewis, "Intelligent selection of language model training data," in *Proceedings of the ACL 2010 Conference Short Papers*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 220–224.

[24] 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*, ser. EMNLP '09, Singapore, 2009.

[25] J. Niehues and A. Waibel, "An MT Error-Driven Discriminative Word Lexicon using Sentence Structure Features," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, 2013, pp. 512–520.

[26] 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.

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

[28] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009.

[29] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy, "Parallel corpora for medium density languages," in *Recent Advances in Natural Language Processing (RANLP 2005)*, 2005, pp. 590–596.

[30] 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.