The KIT Translation Systems for IWSLT 2015

Thanh-Le Ha, Jan Niehues, Eunah Cho, Mohammed Mediani and Alex Waibel

Institute for Anthropomatics and Robotics
KIT - Karlsruhe Institute of Technology, Germany
firstname.lastname@kit.edu

Abstract

In this paper, we present the KIT systems participating in the TED translation tasks of the IWSLT 2015 machine translation evaluation. We submitted phrase-based translation systems for three directions, namely English→German, German→English, and English→Vietnamese. For the official directions (English→German and German→English), we built systems both for the machine translation (MT) as well as the spoken language translation (SLT) tracks.

This year we improved our systems' performance over last year through n-best list rescoring using neural network-based translation and language models and novel discriminative models based on different source-side features and classification methods.

For the SLT tracks, we used a monolingual translation system to translate the lowercased ASR hypotheses with all punctuation stripped to truecased, punctuated output as a preprocessing step to our usual translation system. In addition to punctuation insertion, we also trained that system for sentence boundary insertion since the SLT's data this year come with no sentence boundary.

1. Introduction

The Karlsruhe Institute of Technology participated in the IWSLT 2015 Evaluation Campaign with systems for English→German, German→English and English→Vietnamese. All systems were submitted for the machine translation track, with additional systems for the spoken language translation track in the official directions (English→German, German→English). This year we participated to the new translation direction: English→Vietnamese and we also conducted a short investigation on the impact of word segmentation in our MT system.

On the translation tasks, we integrated new discriminative word lexicon (DWL) models (section 4). We also featured an innovative rescoring method which allows us to take the whole n-best list into account and scale our systems to many features (section 6). Using this, we could seemlessly integrate plentiful numbers of features including the features from the same category, for examples, different DWL models or different neural network language models (section 5).

For SLT tasks, the handling of ASR input was further refined with sentence boundary insertion using a monolingual translation system called *MonoTrans* (section 3). The *MonoTrans* outperformed the provided baseline system for sentence segmentation.

Our baseline system for all translation tasks will be described in section 2. Following sections will present the focused points of this year's KIT systems. After that, the results of the different experiments for the official MT tasks as well as our English→Vietnamese translation will be reported in details in Section 7, before we summarize our findings in Section 8.

2. Baseline system

Our translation systems were conducted using our in-house phrase-based decoder [1]. In English→German and German→English directions, the parallel sections of TED, EPPS, NC and Common Craw are used while TED is the only corpus that we employed to build the English→ Vietnamese system. Addition to the monolingual parts of those corpora, the English News Discussions and Gigaword data are also included in training German→English language models.

The data is preprocessed prior to training and translation. Exceedingly long sentences and aligned sentence pairs having a big difference in length are removed. Special dates, numbers and symbols are normalized. Smartcasing are applied as well. Compound splitting is also conducted to German source texts following the suggestions of [2]. Word segmentation and other typical preprocessing steps for our English→Vietnamese translation system are investigated in details. In addition, our preprocessing also assure that not all sentences from the corpora are used. The noisy ones from Common Crawl were filtered out by a trained SVM classifier as described in [3].

After preprocessing, GIZA++ Toolkit [4] is utilized to perform word alignments over the parallel data. The alignments are then combined to build the phrase table using Moses toolkit [5]. We use the approach described in [6] to adapt out-of-domain phrase tables into the in-domain phrase table from TED for English→German and German→English systems while no adaptation is applied to the English→Vietnamese one.

In both English→German and German→English systems, 4-gram language models with modified Kneser-Ney

smoothing were trained using the SRILM toolkit [7] and scored in the decoding process using KenLM [8]. For English→Vietnamese direction, a longer context of six words is featured in training and scoring.

In addition to conventional word-based language models, we used other language models which are not based on words but contextual information of words. The bilingual language model, based on a four consecutive pairs of source and target words, is used to increase the bilingual context during translation beyond phrase boundaries as described in [9]. On the other hand, the Part-of-Speech (POS) based language model utilizes morphological information by considering a 9-gram sequence of POS tags. Furthermore, we also used the cluster language model based on series of word classes induced by the MKCLS algorithm [10]. This helps alleviate the sparsity problem of surface words by replacing every word in the training corpus with its word class ID.

In our translation systems, we employ two types of reordering models. The first one performs pre-reoderings on the source side by applying the reordering rules learned from POS information [11, 12] and tree constituents [13]. The POS sequences tagged by TreeTagger [14] are used to produce short- and long-range reordering rules. The parsed trees produced by Stanford Parser [15, 16] are used to perform tree-based reorderings which are proved to be helpful for long-dependency modeling. The resulting reordering possibilities for each source sentence are then encoded in a lattice. The second type is the lexicalized reordering model [17] which stores reordering probabilities for each phrase pair scored from the phrase table and the word alignments produced in previous phases.

Other models, described further in following sections, are integrated into our log-linear framework as features. The corresponding weights of those features are tuned using Minimum Error Rate Training (MERT) against the BLEU score as described in [18].

Some additional features, such as source DWL, neural network-based DWL, neural network-based translation and language models, are incorporated into our systems via the ListNet-based rescoring scheme. We will explain further those features as well as our new rescoring approach later in this paper.

3. Preprocessing for speech translation

Since conventional automatic speech recognition (ASR) systems generate either no or only unreliable punctuation marks and sentence segmentation, we design an additional preprocessing step for the test sets of SLT task. In this step, punctuation marks, segmentation, and case information are augmented using a monolingual translation system [19].

Recently, monolingual translation system has shown good performance in inserting punctuation marks for translating speech data [20, 21]. The importance of having proper sentence boundaries, especially, is more emphasized in the IWSLT evaluation campaign 2015. Unlike the SLT condition of previous years' evaluation campaigns, no sentence boundaries are available. Therefore, we need a system which inserts punctuation marks as well as reliable sentence boundaries.

Following previous research described in [22], we built a monolingual translation system which can also augment sentence boundaries. This preprocsesing will be denoted as *MonoTrans*. We built the *MonoTrans* systems for English and German and applied them to two official SLT tracks, English→German and German→English.

For building the systems, we took the preprocessed source side of the parallel training data (either English→German or German→English) and removed the original sentence boundaries. Instead, we inserted sentence boundaries randomly. Therefore, the models can observe sentence boundaries in various positions. If we use the original corpus as it is, the models will learn to insert a sentence boundary at the end of each sentence. This corpus will serve as the target side data of our *MonoTrans* systems.

In order to create the source side data of the *MonoTrans* systems, we remove all punctuation marks from the data and lowercased all words.

Test data is prepared differently using the shifting window of 10 as described in [22]. In this way, each word can be observed in various contexts. Depending on how often a certain punctuation mark was followed by each word, it is inserted based on an empirically chosen threshold.

For both English and German input data, we used the same models in the *MonoTrans* systems. For training data, we used Europarl, TED, NC, and noise-filtered common crawl data, which sums up to 107 million words for English and 85 million words for German. The alignment between non-punctuated, lower-cased text and punctuated, cased text is obtained from GIZA++ [4].

We used a 4-gram language model built on the entire punctuated data using the SRILM Toolkit [7]. In addition to a bilingual language model [9], a 9-gram part-of-speech-based language model is used. The POS is learned from TreeTagger [14]. Also, a $1,000$-class cluster is trained on the punctuated data. The cluster codes are then used to built the additional 9-gram language model. The models were optimized on the official test set of IWSLT evaluation campaign in 2012.

4. Discriminative Word Lexicon

Discriminative Word Lexicon was first introduced by [23]. DWL estimates the probability of a target word appearing in the translation given the source sentence's words. In the original work, a maximum entropy (MaxEnt) model is trained for every target word to determine whether it should be in the translated sentence or not using one feature per source word.

In [24], the authors extended this conventional DWL with n-gram source and target context features. In this evaluation campaign, however, we use the source context features only since the target context features do not bring any im-

provements in our final system. The model using source context features will be referred to as source-context DWL. The source sentence is represented as a bag-of-n-grams, instead of a bag-of-words. This allows us to include local information about source word order in the model.

In addition to this DWL, we integrated a DWL in the reverse direction in rescoring. We will refer to this model as source DWL. This model predicts the target word for a given source word as described in details in [25].

In a first step, we identify the 20 most frequent translations of each word. Then we build a multi-class classifier to predict the correct translation. For the classifier, we used a binary maximum-entropy classifier[1] trained using the one-against-all approach.

As features for the classifier, we used the previous and following three words. Each word is represented by a continuous vector of 100 dimensions as described in [26].

Using the predictions, we calculated two additional features. The first feature is the absolute number of words, where the translation predicted by the classifier and the translation in the hypothesis is the same. The second feature is the sum of the word to word translation probabilities predicted by the classifier that occur in the hypothesis.

While those DWL models can improve the translation by using local source contexts, they employ MaxEnt classifiers which are linear. Hence, they could not really discriminate well the dependencies among features, e.g. a bigram contains two unigrams which somehow reflect a similar or related semantic feature. On the contrary, non-linear classifiers can model those dependencies better since they have the ability to learn some distinct features on higher abstraction levels. [27] introduces non-linearity into DWL by using a deep architecture of neural networks as the alternative classifier. This is referred as neural network-based Discriminative Word Lexicon (NNDWL) in our system. Furthermore, instead of building an indenpendent MaxEnt model for every target word, using NNDWL could improve the translation because it can be seen as a multi-variate classifier consisting of many classifiers which share information among source and target words.

All the DWL models are trained on TED corpus. As showned in previous work, there is no significant improvement using the DWL models trained on bigger corpora.

5. Neural Network Language Model

The traditional n-gram language model (LM) has been applied successfully in many areas of Natural Language Processing due to its robust and simple principles. However, there are some disadvantages of n-gram LM preventing it to better model the cohesion of texts. One of these disadvantages is that the n-grams are presented in a discrete space, hence, it would be hard to estimate well the probability of unseen n-grams which are semantically related to the n-grams appeared in the training set. Continuous space language models, such as restricted boltzmann machine-based LMs[28] or neural network LMs, have been introduced to solve this problem. Basically, in a neural network LM, the discrete representation of words is linearly transformed to a multi-dimentional continuous space. Then one or two following non-linear hidden layers and a softmax output layer are in charge of the probability estimation of the current word based on the transformed representation of the previous words. The transformation and estimation are jointly learned during training. To reduce the time-consuming calculation of the softmax layer, some advanced structures of the output layer and better training methods are proposed[29, 30].

We experimented with different neural network language model toolkits. We used the Torch framework[2], referred to as NNLM, and the nplm toolkit[3][31], referred to as NPLM, to train a feed forward language model. We used in both cases a context of $n = 8$ and trained the model only on the TED corpus. The scores of those language models were added to the n-best list.

6. ListNet-based MT Rescoring

In order to facilitate more complex models, such as the aforementioned DWL models or the neural network language models, we need some way to integrate them to the baseline scores of the phrase-based system. The natural approach is that we rescored the n-best list of candiates in order to select better translations. Compared to other rescoring methods, we would prefer to take the whole list instead of one or two best candidates, so we implemented the rescorer using the ListNet algorithm [32, 33].

This technique defines a probability distribution on the permutations of the list based on the scores of the log-linear model and one based on a reference metric. Therefore, a sentence-based translation quality metric is necessary. In our experiments we used the BLEU+1 score introduced by [34]. Then the rescoring model was trained by minimizing the cross entropy between both distributions on the development data.

Using this loss function, we can compute the gradient with respect to the weight ω_k as follows:

$$\Delta\omega_k = \sum_{j=1}^{n^{(i)}} f_k(x_j^{(i)}) * \quad (1)$$
$$\left(\frac{\exp(f_\omega(x_j^{(i)}))}{\sum_{j'=1}^{n^{(i)}} \exp(f_\omega(x_{j'}^{(i)}))}\right.$$
$$\left. - \frac{\exp(BLEU(x_j^{(i)}))}{\sum_{j'=1}^{n^i} \exp(BLEU(x_{j'}^{(i)}))}\right)$$

When using the i^{th} sentence, we calculate the derivation by

[1] http://hal3.name/megam/

[2] http://torch.ch/

[3] http://nlg.isi.edu/software/nplm/

summing over all $n^{(i)}$ items of the k-best lists. The k^{th} feature value $f_k(x_j^{(i)})$ is multiplied with the difference. This difference depends on $f_\omega(x_j^{(i)})$, the score of the log-linear model for the j hypothesis of the list and the BLEU score $BLEU(x_j^{(i)})$ assigned to this item. Using this derivation, we used stochastic gradient descent to train the model. We used batch updates with ten samples and tuned the learning rate on the development data. The training process ends after 100k batches and the final model is selected according to its performance on the development data.

The range of the scores of the different models may greatly differ and many of these values are negative numbers with high absolute value since they are computed as the logarithm of relatively small probabilities. Therefore, we normalized all scores observed on the development data to the range of $[-1, 1]$ prior to rescoring.

7. Results

In this section, we present a summary of our experiments for all MT and SLT tasks we have carried out for the IWSLT 2015 evaluation. All the reported scores are case-sensitive BLEU scores calculated based on the provided development and test sets.

7.1. German→English

System	MT		SLT
	Dev	Test	Test
Baseline	26.91	28.69	16.57
+ MKCLS	26.97	29.39	-
+ DWL	27.16	29.67	-
KB Mira Rescoring	26.34	29.61	-
+ sDWL + NNDWL	-	**29.91**	16.89

Table 1: Experiments for German→English (MT)

Table 1 presents the results of our experiments for German→English. `tst2012` and `tst2013` are the development and test sets published by the evaluation organizers. Our baseline system already incorporated a number of advanced models. Reorderings were done using both pre-ordering rules as well as a lexicalized reordering model. We adapted the in-domain and out-of-domain phrase tables using the union candidate selection method. In addition to the large language model trained on all available English data, our baseline used an in-domain language model. A bilingual language model trained on all paralled data was also included in the baseline. When we added a 9-gram in-domain cluster language model trained with 100 word classes, our German→English system gained a 0.7 BLEU point improvement. Using a conventional DWL trained on the in-domain data brought further improvement of almost 0.3 BLEU score. The system at this time was used to produce a list of 300 best translation candidates prepared for rescoring. We tried our rescoring using different strategies such as MERT, PRO, KB Mira and ListNet. The corresponding results on a validation set helped us to choose KB Mira as the best strategy to perform rescoring in this direction. Using this strategy, we rescored the n-best list using the old features and two DWL features from source DWLs (sDWL) and neural network-based DWLs (NNDWL). This achieved our best system with 0.3 BLEU points better than the previous system.

For the spoken language translation tasks, since this year's evaluation does not provide the sentence boundaries, we applied the monolingual translation system for sentence boundary and punctuation insertion as well as smart casing described in the section 3. As a baseline for the task, we used our baseline system from the MT task to translate the SLT texts which are already applied *MonoTrans*. Testing on `tst2013` (after removing all sentence boundaries, puntuations and casing), we got the BLEU scores of 16.57. When we applied our best-performing system from the MT task, the SLT system gained an improvements of 0.32 BLEU scores. We submitted this system as our primary system for German→English SLT task. This system achieved 19.64 BLEU score on the official test set this year (`tst2015`). To show the impact of our sentence boundary and punctuation insertion *MonoTrans*, we also submitted another system as the contrastive one. It is the result that we used our best MT system to translate the official SLT test set in which sentence boundaries and punctuations had been inserted by a baseline system provided by the organizer. This contrastive system has a score of 11.84 BLEU points, 7.8 BLEU points less than our primary system on `tst2015`.

7.2. English→German

We conducted several experiments for English→German translation. They are summarized in Table 2. The development set is the `tst2012` and the test set is the `tst2013` data published by the evaluation organizers. The baseline translation system is a phrase-based translation system using two reordering models mentioned above. The phrase table is adapted from the out-of-domain to in-domain TED data. Word-based and non-word language models such as bilingual, POS-based and cluster language models are integrated in the system. Conventional DWLs using source n-grams are also utilized in this phase. The baseline was tuned by MERT and achieved 25.07 and 26.21 BLEU points for development and test sets, respectively.

We performed the rescoring using the ListNet algorithm described in Section 6 on the n-best translation candidates produced by the baseline system. The features that we used are the scores from source and neural network-based DWL models, as well as the neural network-based language models. Adding source DWLs in rescoring scheme helped to improve the system by around 0.7 BLEU points. The NNDWL gained almost 0.2 BLEU points more. Finally, the neural network-based language models, NNLM and NPLM, in-

creased the performance of our system for more than 0.3 BLEU points, reaching 27.50 BLEU points. This system was submitted as our primary system for English→German.

System	Dev	Test
Baseline	25.07	26.21
ListNet Rescoring	24.27	26.36
+ sDWL	-	26.90
+ NNDWL	-	27.18
+ NNLM + NPLM	-	**27.50**

Table 2: Experiments for English→German (MT)

We participated in the spoken language translation tasks for English→German by translating the output of *Mono-Trans* using our best system in the MT task. We got a score of 16.18 BLEU points on the SLT task's official test set `tst2015`.

7.3. English→Vietnamese

This year the IWSLT evaluation organizers have introduced English→Vietnamese translation task for the first time. From the MT perspective, there are two main problems when translating English to Vietnamese: First, the own characteristics of an analytic language like Vietnamese make the translation harder. Second, the lack of Vietnamese-related resources as well as good linguistic processing tools for Vietnamese also affects to the translation quality.

Vietnamese is an analytic language[4]. There are no inflectional morpheme and only several derivational morphemes. In the contrary, it uses a wide variety of function words, temporal or numerical expressions to reflect the grammatical changes. In the linguistic aspect, we might consider Vietnamese is a morphological-poor language, compared to English, German, Finnish or Arabic. In reality, however, the rich set of pronouns in Vietnamese makes the translation to the language harder.

Another linguistic problem which increases the difficulty of Vietnamese-related translation tasks is that the main word boundary marker in Vietnamese is not white space. White spaces are used to separate syllables in Vietnamese, not words. A Vietnamese word consist of one or more syllables. Thus, like Chinese, Vietnamese text processing tools have to deal with ***Word Segmentation*** problem, i.e. how to determine the word boundaries in Vietnamese texts. Word Segmentation is often the first step to be done in a preprocessing phase in those tools since the basic unit is word, not syllable. In this campaign, we also conducted a short investigation to show the importance of using word segmentation methods in an MT system. It would be helpful for further research work on building such translation systems.

Table 3 shows the development stages of the English→Vietnamese system trained on word-segmented texts. We used vnTokenizer[5] [35] for word segmentation and tokenization. The weights of our phrase-based system were also optimized using MERT on word-segmented texts of `tst2012`. And the reported scores were the BLEU scores when we tested the system on word-segmented `tst2013`.

System	Dev	Test
Baseline	19.04	19.97
+ Prereordering	19.87	20.93
+ BiLM + mkcls	20.03	21.07
+ DWL	20.40	21.42

Table 3: Experiments for English→Vietnamese (MT)

The prereordering using POS-based and Tree-based rules helped the most, improving more than 0.8 BLEU points on the development set and nearly 1.0 BLEU points on the test set. This result was not surprising since Vietnamese and English have large differences in term of word order. Integrating non-word language models, e.g bilingual and cluster LMs, brought slightly improvements on both development and test sets, which were 0.16 and 0.14 BLEU points, respectively. In addition, the system gained further enhancement of 0.35 BLEU scores on the test data when we used source-context DWLs. This was the final system we submitted as the primary to the evaluation.

7.3.1. Word-segmented vs. No word-segmented

To compare our methods trained on word-segmented texts and the texts without word segmentation, we built similar systems trained on those two versions and tested them on a non-segmented independent test set. Table 7.3.1 reports the differences. The Dev* and Test* are the BLEU scores measured on the word-segmented development and test sets, respectively. The others are measured on non-segmented datasets.

On the non-segmentation version, we observed that adding more models into the system always helps. And the effects of the models were quite similar to what we observed in case of word-segmented version. For example, the POS- and tree-based reorderings gained the best improvements and integrating DWL were helpful as well as adding non-word language models. The only exception happened when we conducted lexicalized reordering on the word-segmented version, we noticed a slight degrading in the BLEU scores.

It is interesting to observe that our system trained on the unsegmented version of texts performed better than the one trained on the word-segmented texts. One reason we might use to explain this observation is that the vietnamese word segmentation tool, vnTokenizer, is not good enough for TED data. While it simply brings longer contexts, its quality might

[4] `https://en.wikipedia.org/wiki/Vietnamese_language`

[5] `http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer`

System	No Word Segmentation		Word Segmentation		
	Dev	Test	Dev*	Test*	Test
Baseline	24.65	25.66	19.04	19.97	24.95
+ Prereordering	25.55	26.58	19.87	20.93	25.95
+ BiLM	25.58	26.76	19.89	20.99	26.36
+ mkcls	25.77	26.85	20.20	21.12	26.43
+ DWL	25.83	27.18	20.40	**21.42**	26.55
+ Lexicalized Reordering	**25.99**	**27.64**	**20.41**	21.24	**26.62**

Table 4: Experiments for English→Vietnamese

affect the word alignments, which in turn affect to other components in our system. In addtion, the advantages of using longer context in case of training on word-segmented texts can be covered somehow by phrase extraction and language modeling. Since phrases in our MT are basically sequences of words, we can see a phrase in the non-segmented system as a shorter phrase compared to corresponding one in the word-segmented system. We would need a more comprehensive investigation on this problem. Due to the fact that we have been investigating the unsegmented system after the submission deadline, we did not submit the system despite its better performance.

8. Conclusions

In this paper, we described several innovative works that we applied to our translation systems we participated in the IWSLT 2015 Evaluation Campaign. Besides the traditional, official MT and SLT tasks for English→German and German→English, we also submitted the newly published translation tasks English→Vietnamese.

For all official translation directions, we built strong baseline systems including our advanced reordering methods, data selection and adaptation techniques, as well as several word-based and non-word language models. Those individual models proved successful in many of the systems.

The notable enhancement this year is the n-best list rescoring which performed better than other MT optimization techniques and scaled better to a large number of features. We used this rescoring to leverage newly-added features such as the DWLs or other neural language models.

The combination of new features with the traditional features in a rescoring scheme boosted our translation systems in both English→German and German→English direction to more than 1.2 BLEU points improvements. When we applied our techniques for English→Vietnamese, we observed the improvements brought by the individual components. We also showed the effects of using non word-segmented texts in training such a translation system.

A monolingual translation system for punctuation insertion played a vital role in adjusting the ASR output for speech translation. This system was also capable to perform decent sentence segmentation which is necessary for the SLT data this year when they do not have any sentence boundary.

9. Acknowledgements

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 645452.

10. References

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

[2] P. Koehn and K. Knight, "Empirical Methods for Compound Splitting," in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, 2003.

[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 8th International Workshop on Spoken Language Translation*, San Francisco, CA, USA.

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

[5] 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 the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 2007.

[6] J. Niehues and A. Waibel, "Detailed Analysis of Different Strategies for Phrase Table Adaptation in SMT," in *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas*, San Diego, CA, USA, 2012.

[7] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit." in *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO, USA, 2002.

[8] K. Heafield, "KenLM: faster and smaller language model queries," in *Proceedings of the 6th Workshop on Statistical Machine Translation*, Edinburgh, United Kingdom, 2011.

[9] J. Niehues, T. Herrmann, S. Vogel, and A. Waibel, "Wider Context by Using Bilingual Language Models in Machine Translation," in *Proceedings of the 6th Workshop on Statistical Machine Translation*, Edinburgh, United Kingdom, 2011.

[10] F. J. Och, "An Efficient Method for Determining Bilingual Word Classes," in *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway, 1999.

[11] 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*, Skövde, Sweden, 2007.

[12] J. Niehues and M. Kolss, "A POS-Based Model for Long-Range Reorderings in SMT," in *Proceedings of the 4th Workshop on Statistical Machine Translation*, Athens, Greece, 2009.

[13] T. Herrmann, J. Niehues, and A. Waibel, "Combining Word Reordering Methods on different Linguistic Abstraction Levels for Statistical Machine Translation," in *Proceedings of the 7th Workshop on Syntax, Semantics and Structure in Statistical Translation*, Altanta, GA, USA, 2013.

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

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

[16] D. Klein and C. D. Manning, "Accurate Unlexicalized Parsing," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 2003.

[17] 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 2nd International Workshop on Spoken Language Translation*, Pittsburgh, PA, USA, 2005.

[18] A. Venugopal, A. Zollman, and A. Waibel, "Training and Evaluation Error Minimization Rules for Statistical Machine Translation," in *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, MI, USA, 2005.

[19] S. Peitz, M. Freitag, A. Mauser, and H. Ney, "Modeling punctuation prediction as machine translation."

[20] T.-L. Ha, J. Niehues, T. Herrmann, M. Mediani, E. Cho, Y. Zhang, I. Slawik, and A. Waibel, "The KIT Translation Systems for IWSLT 2013," in *Proceedings of the International Workshop on Spoken Language Translation*, ser. IWSLT 2013, Heidelberg, Germany, 2013.

[21] I. Slawik, M. Mediani, J. Niehues, Y. Zhang, E. Cho, T. Herrmann, T.-L. Ha, and A. Waibel, "The KIT Translation Systems for IWSLT 2014," in *Proceedings of the International Workshop for Spoken Language Translation (IWSLT 2014)*, Lake Tahoe, CA, USA, 2014.

[22] E. Cho, J. Niehues, and A. Waibel, "Segmentation and Punctuation Prediction in Speech Language Translation using a Monolingual Translation System," in *Proceedings of the 9th International Workshop on Spoken Language Translation*, Hong Kong, 2012.

[23] 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.

[24] J. Niehues and A. Waibel, "An MT Error-Driven Discriminative Word Lexicon using Sentence Structure Features," in *Proceedings of the 8th Workshop on Statistical Machine Translation*, Sofia, Bulgaria, 2013.

[25] , "Source Discriminative Word Lexicon for Translation Disambiguation," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT15)*, Danang, Vietnam, 2015.

[26] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations." in *HLT-NAACL*, 2013, pp. 746–751.

[27] T.-L. Ha, J. Niehues, and A. Waibel, "Lexical Translation Model Using a Deep Neural Network Architecture," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT14)*, Lake Tahoe, CA, USA.

[28] J. Niehues and A. Waibel, "Continuous space language models using restricted boltzmann machines." in *IWSLT*, 2012, pp. 164–170.

[29] H.-S. Le, I. Oparin, A. Allauzen, J.-L. Gauvain, and F. Yvon, "Structured output layer neural network language model," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5524–5527.

[30] H. Schwenk, A. Rousseau, and M. Attik, "Large, pruned or continuous space language models on a gpu for statistical machine translation," in *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*. Association for Computational Linguistics, 2012, pp. 11–19.

[31] A. Vaswani, Y. Zhao, V. Fossum, and D. Chiang, "Decoding with Large-Scale Neural Language Models Improves Translation." in *EMNLP*, 2013, pp. 1387–1392.

[32] J. Niehues, Q. K. Do, A. Allauzen, and A. Waibel, "Listnet-based MT Rescoring," *EMNLP 2015*, p. 248, 2015.

[33] Z. Cao, T. Qin, T. yan Liu, M.-F. Tsai, and H. Li, "Learning to Rank: From Pairwise Approach to Listwise Approach," in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, USA, 2007, pp. 129–136.

[34] P. Liang, A. Bouchard-Côté, D. Klein, and B. Taskar, "An End-to-end Discriminative Approach to Machine Translation," in *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006)*, Sydney, Australia, 2006, pp. 761–768.

[35] H. P. Le, T. M. H. Nguyen, R. Azim, and T. V. Ho, "A Hybrid Approach to Word Segmentation of Vietnamese Texts," *Language and Automata Theory and Applications*, pp. 240–249, 2008.