Source Discriminative Word Lexicon for Translation Disambiguation

Teresa Herrmann, Jan Niehues, Alex Waibel

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

Abstract

This paper presents a source discriminative word lexicon that performs translation disambiguation for individual source words using structural features, such as context and dependency relations in the sentence. The individual translation predictions are combined into sentence scores that are used in N-best list re-ranking to improve the translation output of a state of the art phrase-based machine translation system. The approach is used to improve explicitly the translation of word categories that require grammatical agreement to hold in the target language after translation, e.g. pronouns, as well as subjects and verbs. The results show that the translation predictions provided by the source discriminative word lexicon increase the prediction accuracy by up to 10%. The translation quality can be improved by up to 0.6 BLEU points on English-German translation.

1. Introduction

Ambiguity of words is a big challenge for all natural language processing tasks. Already within the same language, words can be ambiguous with regard to their part-of-speech (***can**, n. - **can**, v.*), word sense (***bank**, n., financial institution - **bank**, n., side of a river*) or what they are referring to in the given context (*The monkey eats the banana. **It** is brown.*). For translation, such ambiguities pose an additional difficulty. Unless the very same ambiguity exists in the target language, the ambiguity needs to be resolved in order to generate the correct translation. When translating into German, for example, depending on the correct part-of-speech, word sense and antecedent in the sentence, the translation for each of those examples is a different one.

The word(s) indicating which is the correct word sense or antecedent for an ambiguous word in a given context, could occur in a more distant part of the sentence. That means long-range dependencies need to be considered in order to generate the correct translation. We propose a discriminative framework for modeling these dependencies utilizing any conceivable set of features for predicting the correct translation. We show the potential of this approach in detail on the third type of ambiguity mentioned above: The translation of pronouns, which is conditioned on the translation of the antecedent they refer to, since the pronoun in the target language needs to share the morphological properties of the antecedent in the target language.

An approach to explicitly performing anaphora resolution to uncover the pronoun-antecedent relationship for pronoun translation disambiguation was carried out in [1]. Their experiments motivated the present work, however the approach was adapted in the following ways: Instead of focusing only on third person pronouns, we include all personal pronouns and also take translations into other word categories into account. In order to allow for a more comprehensive exploration of the source discriminative word lexicon approach we apply it for translation disambiguation for all words and perform separate evaluation of the performance on pronouns. We further evaluate it on another difficult agreement task, the agreement of subject and verb in a sentence.

State-of-the-art machine translation systems struggle with these particular kinds of linguistic requirements [2]. Hence, we believe our approach can provide a comprehensive solution for many of these challenges where long-range dependencies have to be met in order to ensure congruency of linguistic features.

2. Related Work

Already one of the early statistical approaches performs word sense disambiguation by defining senses according to the different translations of a word [3]. Since then, several approaches integrate word sense disambiguation into phrase-based [4] or hierarchical [5] translation systems or use it in N-best list re-ranking [6]. Context features as well as dependencies have been used to perform word sense disambiguation for different close and distant language pairs [7, 8].

Apart from applying actual word sense disambiguation in machine translation, linguistic information, such as context words, dependencies or syntax can be integrated in machine translation as additional features in order to improve the translation quality [9, 10].

Among the approaches that particularly model translation prediction as is done in this paper, [11] predict the occurrence of a target word in a translated sentence given the source words using a discriminative approach. Similar approaches operating use a multilayer perceptron [12] or a bilingual neural network to learn abstract word representations and features in order to predict word, stem and suffix translations for source words given the source context [13].

An approach that integrates discriminative classifier predictions based on context and POS tags into decoding is presented in [14].

There is limited research on modeling anaphora resolution for the translation of pronouns in a statistical machine translation system. The first approaches integrate the output of explicit anaphora resolution components within the MT system [15] often focusing in particular on neuter pronouns [16, 17, 18] with limited improvements. [19] perform a classification using automatic anaphora resolution output, discriminating between the possible French pronouns in the translation. Their neural network approach surpasses maximum entropy classification and can even be extended to perform latent anaphora resolution and translation prediction jointly. An approach for translation of the English pronoun *it* into Czech is modeled by classification of the pronoun into one of three classes triggering different treatment in the tree-to-tree-based machine translation system [20].

Since the success of machine translation depends to a great deal on the morphological complexity of the target language [21], modeling target morphology in various ways is a popular direction of research. There are many approaches that perform a prediction of the inflected word forms in the target language. Conditional random fields are a popular approach to sequence labeling which is applied to predict morphemes [22], morphosyntactic properties [23], or inflection [24, 25, 26] on the target side. The latter perform a two-step translation process, first translating into stemmed forms or lemmas and then predicting the fully inflected forms. Two-step translation into Czech [27] applies two translation systems sequentially, to translate first into simplified Czech and then into fully inflected Czech. Factored translation models treat word, lemma, part-of-speech and morphological features as separate factors and perform morphological generation in a phrase-based machine translation system [28]. Enriching the source language with linguistic information in order to address noun phrase and subject-verb agreement [29]. and using fixed-length suffixes in order to improve grammaticality of the translation output [30] are further applications of the factored model.

The presented approach is modeled based on the idea of the discriminative word lexicon [11, 31], however operating on the source side instead of the target side and predicting translations given source side features. In contrast to approaches operating on the phrase level, we model predictions for individual words, however taking a up to 6 neighboring words into account and therefore covering longer context than included in the average phrase length. The approach is closely related to [14], but differs by modeling predictions for words instead of phrases, which are less sparse and therefore should provide better estimates. In addition, we include dependency features which can cover longer distances and more implicit dependencies in the sentence.

3. Source Discriminative Word Lexicon

We implement the translation disambiguation as a prediction task. The prediction is motivated by the discriminative word lexicon [31]. While the discriminative word lexicon (DWL) operates on the target side and learns to predict for each target word whether it should occur in a given target sentence, the source discriminative word lexicon (SDWL) operates on the source side. For every source word a classifier is trained to predict its translation in the given sentence. We perform a multi-class classification task by identifying for every source word the 20 most frequent translations as provided by the word alignment generated with GIZA++. All target language words that occur less often than the 20 most frequent words are assigned to one class, called **other**. Alignments to the NULL word on the target side are treated in the same way as if NULL were a word. We limit the source vocabulary to the words occurring in the test data and train up to 20 classifiers for each source word. In reality, most words have a lot less than 20 alternative translation options. The SDWL uses binary maximum entropy classifiers trained using the one-against-all scheme. That means we use a maximum entropy model to estimate $p(e|f, c(f))$, where e is the target word we want to predict given source word f and its context/dependency features $c(f)$. During training the maximum entropy models for the individual classes for each source word are learned based on the given set of features extracted from the source sentence and the correct class of each training example. For the prediction, the test data is first separated into words. For each word the features are extracted from the source sentence it stems from. Then all the binary maximum entropy models for the multiple classes are applied and each of them produces a prediction. The final prediction corresponds to the class with the highest prediction probability.

3.1. Structural Features

The training examples and test data for the classifiers are represented by a set of features and the class this example belongs to. We experiment with different types of features representing the structure of a sentence to varying degrees.

3.1.1. Bag-of-Words

A straight forward way to represent the source sentence for this classification task is to use the bag-of-words approach. This is the least structural informative feature which does not provide any knowledge about the sentence beyond the mere existence of the words in it.

3.1.2. Context

The context feature adds structural information about the local context of the modeled source word in the sentence. In addition to the context words themselves, their position is encoded in the feature such that the same word occurring at

a different position (relative to the source word in question) would result in a different feature. We include up to six context words, three on each side of the source word. Hence, this feature type provides structural information by means of sequential order within a limited context.

3.1.3. Dependency Relations

The feature contributing the most information about the sentence structure is based on the relations between the source sentence words in a dependency tree. In order to obtain the dependency relations, we extract a dependency tree from a constituency parse tree using the Stanford Parser [32, 33]. Then we include the dependency relations between the source word and its parent and children in the dependency tree as features. That means, we form a feature consisting of the governance relation (parent or child of the source word), the dependency relation type (from the set of dependency relations described in [34] e. g., nsubj, dobj, vmod, ...) and the connected word itself. This type of feature allows to capture structure by means of semantic dependencies that can range over longer distances in the sentence, but are relevant due to the semantic connection to the current source word. An example for the features for the word *it* in a given sentence is presented in Example 3.1.

Sentence:		*Well it obviously is not.*
Features:	bag-of-words	not is it obviously well .
	context	-1_well $+1$_obviously $+2$_is
	dependency	dep_parent_nsubj_is

Example 3.1: *Representation of the source word "it" by the different features*

3.2. Word Representation

We compare two methods to represent the words in the features: word IDs and word vectors.

3.2.1. Word IDs

When representing words by word IDs, we use the source vocabulary size V_{source} as the dimension of the feature space, a word's ID in the vocabulary as a feature and we set the feature to 1 if it is used in the example. All other features are set to 0. For accommodating the context features (**context**), we extend the size of the features space such that $V_{context} = c * V_{source}$ where c equals the size of the context. Each position of a word in the context hence has its own range in the features space, and words in different context positions can be distinguished accordingly. The features representing dependency relations (**dep**) are included in a similar fashion. Again, a new feature space is defined as $V_{dep} = d * V_{source}$ where d equals the amount of all dependency relations, where parent and child relations are counted separately. The feature types can be combined by simply concatenating the individual feature spaces. That means when all three types of features are used the size of the feature space amounts to $V_{source} + V_{context} + V_{dep}$. It is obvious, that with this strategy for design the feature space grows quite big, possibly leading to data sparseness problems. In order to reduce dimensions, the representation via word vectors seemed an appropriate measure.

3.2.2. Word Vectors

The word vectors for feature representation are generated using word2vec [35] with the number of dimensions set to 100. That means each word is represented by a 100-dimensional vector. However, it is not straight forward how multiple words should be expressed in this representation, so that the representation by word vectors is not applied for the bag-of-words features, but only for the context and dependency features. In case of the vector representation of the context features (**contextVec**), each position in the context words receives its own range in the feature space. Hence, the size of the feature space equals to $V_{contextVec} = c * dim$, where c is the context size and dim the dimension of the vector representation. This amounts to a significant reduction compared to $V_{context}$ used in the representation method via word IDs. The feature space for dependency relations using word vectors (**depVec**) equals to $V_{depVec} = d * dim$ with d being the inventory of dependency relations. Compared to V_{dep}, again a huge reduction. In addition to the **depVec** feature, further variants of the dependency feature are compared:

parentDepVec
: For this feature, only the dependency relation to the parent word is represented in vector representation.

parentWordVec
: This feature consists of the vector representation of the parent word and an additional binary feature that is 1 if the parent word is the root of the dependency tree.

parentWordVec+DepRel
: In addition to the **parentWordVec** feature, the dependency relation to the parent word is encoded as a vector.

As for the word-based features, word vector features can be combined by concatenation of feature spaces.

3.3. Integration of SDWL Predictions

In order to integrate the individual translation predictions into a machine translation system we use the prediction probabilities for individual words to produce scores for whole sentences. The combination of individual translation predictions for words into a sentence score is explained in the following. These scores are then used in N-best list re-ranking.

3.3.1. SDWL-based Re-ranking Scores

For each of the translation hypotheses in the N-best list, we generate a sentence score based on the translation predictions for the individual words in the sentence. We compare four methods to combine the individual word scores into a sentence score for a particular translation hypothesis.

Absolute number of predicted words
: We count the number of word translations in the sentence that coincide with the predicted translations by the translation prediction model.

Relative number of predicted words
: As an alternative score we again count the number of words in the translation hypothesis that coincide with the predicted translation. This number of matches is then divided by the total number of target words generated by the source words according to the alignment.

Sum of prediction probabilities
: For every source word we sum up the prediction probabilities associated with their aligned words in the hypothesis.

Sum of prediction ranks
: Instead of summing up the prediction probabilities of the words in the hypothesis, we sum up the ranks of the words according to their prediction probability.

All these scores were both used individually and collectively as additional sentence scores for N-best list re-ranking, in order to find out which of them are most beneficial for judging translation quality.

4. Experiments

We perform two types of experiments with the presented source discriminative word lexicon. First we use it independently to predict the translation for individual source words in the sentence and measure the prediction accuracy against the reference translation. Afterwards, we combine the individual predictions for words into a sentence score and use it in N-best list re-ranking of machine translation output.

4.1. Data

We train the classifiers on the parallel training data consisting of TED talks provided for the IWSLT 2014 evaluation campaign. Due to the limitation of the source vocabulary to the test data, we train 26,498 classifiers for 5,389 source words, which equals to an average of 4.9 translation alternatives per word. The prediction accuracy of the source discriminative word lexicon is measured on test2011 and test2012 combined. The impact of the source discriminative word lexicon on translation quality is measured after N-best list re-ranking the output of a machine translation system with the SDWL sentence score. The translation system is tuned on test2011 and tested on test2012. For N-best list re-ranking the three data sets test2010, dev2010, and test2011 are used. Translation quality before and after rescoring is reported on test2012.

4.2. Translation System

The re-ranking experiment is done using a phrase-based machine translation system. The phrase table is built using the Moses toolkit [36] and n-gram language models are trained with the SRILM toolkit [37]. Translations are generated with a phrase-based MT decoder [38]. Optimization is done with a variant of MERT [39]. Translation quality is measured in BLEU [40].

In addition to the basic translation model and language model, the system applies several word-based, POS-based and cluster-based language models, as well as a bilingual language model. Furthermore, an original discriminative word lexicon for the target side is included. Several word reordering models are used. Tree-based and POS-based reordering rules produce reordering variants of each source sentence stored in a word lattice and a lexicalized reordering model provides probability scores for the order of phrases in the translation hypotheses produced by the decoder. The translation system is described in detail in [41]. In addition, the SDWL in reduced form using only bag-of-words and context features is applied in three other systems. A German-English News system [42], an English-German and German-English TED system [43].

4.3. N-best List Re-Ranking

As mentioned above the predictions from the SDWL are combined into sentence scores for the translation hypotheses in the N-best list produced by the translation system. Then N-best list re-ranking is performed as described in [41] using the ListNet algorithm [44].

5. Results

This section presents the results of the translation prediction model tested on English-to-German translation of TED talks. First, we will show that the prediction accuracy improves when applying the proposed set of structural features. In addition, the translation quality can be improved when using the translation predictions for N-best list re-ranking to find a better translation among the hypotheses in the N-best list of the translation system.

5.1. Translation Prediction

We compare the different features for representing the sentence and context for the translation prediction of individual source words described above. We measure the accuracy of the translation prediction achieved with each of the features and feature combinations. Table 1 presents an overview of the experiments. It shows the average prediction accuracy on all words in the data used for testing.

The baseline prediction is performed with a maximum likelihood classifier, which a priori chooses the most frequent class, without using any features at all. We can see that using the bag-of-words features consisting of the words contained in the source sentence already improves over the baseline prediction. When applying the more structurally informative features, both context and dependency features individually improve considerably over the simple bag-of-words features. Among the context feature variants, the vector representation with 2 words of context in both directions performs best. For the dependency features, it is the vector representation using both parent and child relations, which leads to the best predictions. Combining the two best performing features **contextVec** and **depVec**, holds another small improvement leading to a prediction accuracy that is more than 7% higher than the baseline prediction, which corresponds to 14% relative improvement.

	Prediction Accuracy
Baseline	52.09
Bag-of-Words	53.29
Context (+/- 2 words)	58.74
ContextVec (+/- 2 words)	58.97
ContextVec (+/- 3 words)	57.48
Dep	56.07
DepVec	57.27
ParentDepVec	55.02
ParentWordVec	54.65
ParentWordVec+DepRel	55.20
ContextVec (+/-2) + DepVec	59.37

Table 1: *Translation prediction results: all words*

5.1.1. Pronoun Translation

In order to explicitly measure the accuracy of the translation prediction for pronouns, we selected the pronouns among the source words and measured the prediction accuracy of those words. Table 2 presents the prediction accuracy of source language pronouns. The pronouns achieve higher absolute numbers of translation accuracy. However, the improvements by the different types of features is comparable to the improvements on all words. The use of structural features led to an absolute and relative increase in prediction accuracy by more than 5% and 9%, respectively.

5.1.2. Subject-Verb Agreement

We also analyzed the accuracy of prediction features with respect to subject-verb agreement. For this purpose all word pairs connected by a subject relation were extracted from the dependency trees for the source sentences. All words posing as parents in such a dependency relation were taken to be possible verbs, and all children in a subject relation are considered as possible subjects. It has to be noted, though, that

	Prediction Accuracy	
	all words	pronouns
Baseline	52.09	59.58
Bag-of-Words	53.29	60.03
ContextVec (+/- 2 words)	58.97	64.89
DepVec	57.27	63.12
ContextVec (+/-2) + DepVec	59.37	65.08

Table 2: *Translation prediction results: pronouns*

the subject and verb list can also contain words of other parts-of-speech, since relations such as the one between nouns and adjectives can also be defined as a subjective relation in a dependency tree. However, manual inspection confirmed that apart from a few outliers it was indeed mostly words qualifying as subjects and verbs in the extracted list and we chose not to apply an additional manual filter. In order to produce comparable results, we measured the prediction accuracy of the words in the subject and verb lists in the same way as all words and pronouns in the results reported above. The results are presented in Table 3. We can see that the improvements of subjects and verbs are even higher than the ones on pronouns or all words, getting as close as 10% absolute and 20% relative over the baseline prediction.

	Prediction Accuracy		
	all words	subj.	verbs
Baseline	52.09	46.81	46.71
ContextVec (+/-2) + DepVec	59.37	56.00	54.12

Table 3: *Translation prediction results: subjects and verbs*

5.2. N-Best List Re-ranking

The results of improved prediction accuracy of the SDWL model with structural informative features presented above are encouraging. Therefore, we want to use the predictions to judge the quality of a particular translation hypothesis in N-best list re-ranking. For the baseline, an N-best list re-ranking is performed, using the original sentence-based scores available from the translation system. Then we compare the four ways of generating an additional score for a given hypothesis based on the individual word translation predictions described above: absolute and relative number of predicted words in the hypothesis, sum of the prediction probabilities of the words chosen in the hypothesis and rank of the words in the hypothesis according to prediction probabilities. We use the SDWL features that performed best in the previous experiment, i.e. the context vectors with context +/-2 words and the dependency vectors.

Table 4 shows an overview over the results. Three of the methods to create the sentence score perform very similar, providing about 0.2 BLEU points of improvement. Only

Source:	*I memorized in my anatomy class the origins and exertions of every muscle [...]*
Baseline:	*Ich in meinem Anatomie der Klasse die Ursprünge und Strapazen eines jeden Muskel [...] auswendig [...]*
+SDWL:	*Ich in meiner Klasse Anatomie die Ursprünge und Strapazen jeder Muskel [...] auswendig [...]*
Reference:	*In meiner Anatomievorlesung lernte ich die Ursprünge und Ausläufer jedes Muskels [...]*

Example 5.1: *Correct gender for pronoun*

Source:	*There I think that the arts and film can perhaps fill the gap, and simulation.*
Baseline:	*Ich glaube, dass die Kunst und Film kann vielleicht die Lücke füllen, und Simulation.*
+SDWL:	*Ich glaube, dass die Kunst und Film, vielleicht können die Lücke füllen, und Simulation.*
Reference:	*Hier können, denke ich, die Kunst und der Film vielleicht die Lücke füllen, sowie Simulationen.*

Example 5.2: *Correct case agreement between subject and verb*

when using the prediction ranks of the words in the hypothesis, the translation quality is not increased. That means that the translation predictions can indeed serve as an indicator for translation quality when combined in one of the three proposed ways. By using the SDWL-based scores it is possible to select an even better hypothesis from the N-best list compared to using only the available scores from the translation system.

Translation System	TED (2014) EN-DE
Baseline	24.04
SDWL: Abs	24.20
SDWL: Rel	24.22
SDWL: Sum	24.21
SDWL: Rank	23.98

Table 4: *Prediction features in re-ranking: EN-DE TED*

5.2.1. Additional Systems

The SDWL was further applied in several other translation systems in recent evaluation campaigns. Due to time constraints only the context features consisting of +/-3 words were used for the translation prediction. Table 5 shows the improvements that were gained from N-best list re-ranking with the SDWL on German-English translation of News in the WMT 2015 shared translation task as well as German-English and English-German translation of TED talks in the IWSLT 2015 machine translation task. Depending on the language and the task between 0.3 and 0.6 BLEU points can be gained from including the translation predictions even when using only the surrounding 3 context words.

Translation System	News	TED (2015)	
	DE-EN	DE-EN	EN-DE
Baseline	27.87	29.59	26.36
SDWL (ContextVec +/-3)	28.18	29.87	26.90

Table 5: *Prediction features in re-ranking: additional results*

5.3. Translation Examples

Example 5.1 shows an improvement in pronoun translation that was achieved with the SDWL. In this translation the baseline translation produces the pronoun where the gender is incorrect. Within the prepositional phrase the gender of the possessive pronoun needs to agree with its associated noun *Klasse*, which is feminine. When using the SDWL the correct gender is generated in the translation.

Example 5.2 shows that the translation prediction model also encourages morphological agreement between subject and verb. Since the information that the verb is actually in plural form is not encoded in the source language (The English verb *can*, can be both singular and plural), rendering a plural verb in the translation is not straight forward. Hence, the structural features are able to capture the plural subject in the dependency feature and/or the plural indicator *and* in the context feature, and rank the hypothesis higher where the plural verb (*können*) occurs in the translation.

6. Conclusions

We have presented a model for translation disambiguation using structural features in a classification task. The translation of a source word in a given sentence is predicted based on the classification into one of its 20 most frequent translation options. Structural features such as source context words and relations in the dependency tree of the source sentence allow to include knowledge about the sentence structure when modeling the prediction. The model is in particular aimed at improving challenging linguistic issues like the translation of pronouns and generating morphological agreement in the translated sentence.

The prediction results have shown that the accuracy of predicting a translation for individual source words increases considerably when including the context and dependency features. Representing the features by a word2vec word vector representation both reduces dimensions and increases prediction accuracy. Even though the context and dependency features contribute similar improvements individually, their combination provides the highest prediction accuracy. A separate inspection of pronouns, subjects and verbs con-

firms that these were improved in particular by up to 10%.

The individual translation predictions for the source words in each sentence are combined into a sentence score used in N-best list re-ranking. Using the prediction scores in re-ranking provides between 0.2 and 0.6 BLEU points of improvement.

Directions for future work could be the investigation of features that include more semantic information such as the semantic distance between words, or the replacement of the current classification approach by other machine learning techniques such as neural networks which are able to model more implicit dependencies. Furthermore, we would expect a positive effect on the phrase selection, if the predictions were made available already at decoding time.

7. 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.

8. References

[1] J. Weiner, "Pronominal Anaphora in Machine Translation," Master's thesis, Karlsruhe Institute of Technology, Karlsruhe, Germany, 2014.

[2] T. Herrmann, "Linguistic Structure in Statistical Machine Translation," Ph.D. dissertation, Karlsruhe Institute of Technology, Karlsruhe, Germany, 2015.

[3] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "Word-sense Disambiguation Using Statistical Methods," in *Proceedings of ACL 1991*, Berkeley, CA, USA, 1991.

[4] M. Carpuat and D. Wu, "Improving Statistical Machine Translation using Word Sense Disambiguation," in *Proceedings of EMNLP-CoNLL 2007*, Prague, Czech Republic, 2007.

[5] Y. S. Chan, H. T. Ng, and D. Chiang, "Word Sense Disambiguation Improves Statistical Machine Translation," in *Proceedings of ACL 2007*, Prague, Czech Republic, 2007.

[6] L. Specia, B. Sankaran, and M. Graças Volpe Nunes, "n-Best Reranking for the Efficient Integration of Word Sense Disambiguation and Statistical Machine Translation," *Lecture Notes in Computer Science*, vol. 4919, 2008.

[7] A. Max, R. Makhloufik, and P. Langlais, "Explorations in Using Grammatical Dependencies for Contextual Phrase Translation Disambiguation," in *Proceedings of EAMT 2008*, Hamburg, Germany, 2008.

[8] K. Gimpel and N. A. Smith, "Rich Source-side Context for Statistical Machine Translation," in *Proceedings of WMT 2008*, Columbus, OH, USA, 2008.

[9] L. Shen, J. Xu, B. Zhang, S. Matsoukas, and R. Weischedel, "Effective Use of Linguistic and Contextual Information for Statistical Machine Translation," in *Proceedings of EMNLP 2009*, Suntec, Singapore, 2009.

[10] R. Haque, S. K. Naskar, A. van den Bosch, and A. Way, "Integrating Source-language Context into Phrase-based Statistical Machine Translation." *Machine Translation*, vol. 25, no. 3, 2011.

[11] A. Mauser, S. Hasan, and H. Ney, "Extending Statistical Machine Translation with Discriminative and Trigger-based Lexicon Models," in *Proceedings of EMNLP 2009*, Suntec, Singapore, 2009.

[12] A. Patry and P. Langlais, "Prediction of Words in Statistical Machine Translation using a Multilayer Perceptron," in *Proceedings of MT Summit XII*, 2009.

[13] K. M. Tran, A. Bisazza, and C. Monz, "Word Translation Prediction for Morphologically Rich Languages with Bilingual Neural Networks," in *Proceedings of EMNLP 2014*, Doha, Qatar, 2014.

[14] A. Tamchyna, F. Braune, A. M. Fraser, M. Carpuat, H. D. III, and C. Quirk, "Integrating a Discriminative Classifier into Phrase-based and Hierarchical Decoding," *Prague Bull. Math. Linguistics*, vol. 101, pp. 29–42, 2014.

[15] R. Mitkov, S. kwon Choi R, and A. Sharp, "Anaphora resolution in Machine Translation," in *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, 1995.

[16] R. Le Nagard and P. Koehn, "Aiding Pronoun Translation with Co-reference Resolution," in *Proceedings of WMT 2010*, Uppsala, Sweden, 2010.

[17] L. Guillou, "Improving Pronoun Translation for Statistical Machine Translation," in *Proceedings of the Student Research Workshop at EACL 2012*, Avignon, France, 2012.

[18] C. Hardmeier and M. Federico, "Modelling Pronominal Anaphora in Statistical Machine Translation," in *Proceedings of IWSLT 2010*, Paris, France, 2010.

[19] C. Hardmeier, J. Tiedemann, and J. Nivre, "Latent Anaphora Resolution for Cross-Lingual Pronoun Prediction." in *Proceedings of EMNLP 2013*, Seattle, WA, USA, 2013.

[20] M. Novák, A. Nedoluzhko, and Z. Žabokrtský, "Translation of "It" in a Deep Syntax Framework," in *Proceedings of DiscoMT 2013*, Sofia, Bulgaria, 2013.

[21] A. Birch, M. Osborne, and P. Koehn, "Predicting Success in Machine Translation," in *Proceedings of EMNLP 2008*, Honolulu, HI, USA, 2008.

[22] A. Clifton and A. Sarkar, "Combining Morpheme-based Machine Translation with Post-processing Morpheme Prediction," in *Proceedings of ACL-HLT 2011*, Portland, OR, USA, 2011.

[23] S. Green and J. DeNero, "A Class-based Agreement Model for Generating Accurately Inflected Translations," in *Proceedings of ACL 2012*, Jeju, South Korea, 2012.

[24] K. Toutanova, H. Suzuki, and A. Ruopp, "Applying Morphology Generation Models to Machine Translation," in *Proceedings of ACL-HLT 2008*, Columbus, OH, USA, 2008.

[25] A. Fraser, M. Weller, A. Cahill, and F. Cap, "Modeling Inflection and Word-Formation in SMT," in *Proceedings of EACL 2012*, Avignon, France, 2012.

[26] A. E. Kholy and N. Habash, "Translate, Predict or Generate: Modeling Rich Morphology in Statistical Machine Translation," in *Proceedings of EAMT 2012*, Trento, Italy, 2012.

[27] D. Mareček, R. Rosa, P. Galuščáková, and O. Bojar, "Two-step Translation with Grammatical Post-processing," in *Proceedings of WMT 2011*, Edinburgh, Scotland, 2011.

[28] P. Koehn and H. Hoang, "Factored translation models," in *Proceedings of EMNLP-CoNLL 2007*, Prague, Czech Republic, 2007.

[29] E. Avramidis and P. Koehn, "Enriching Morphologically Poor Languages for Statistical Machine Translation," in *Proceedings of ACL-HLT 2008*, Columbus, OH, USA, 2008.

[30] N. S. Razavian and S. Vogel, "Fixed Length Word Suffix for Factored Statistical Machine Translation," in *Proceedings of ACL 2010*, Uppsala, Sweden, 2010.

[31] J. Niehues and A. Waibel, "An MT Error-driven Discriminative Word Lexicon using Sentence Structure Features," in *Proceedings of WMT 2013*, Sofia, Bulgaria, 2013.

[32] D. Klein and C. D. Manning, "Fast Exact Inference with a Factored Model for Natural Language Parsing," in *Proceedings of NIPS 2002*, Vancouver, Canada, 2002.

[33] ——, "Accurate Unlexicalized Parsing," in *Proceedings of ACL 2003*, Sapporo, Japan, 2003.

[34] M.-C. de Marneffe and C. D. Manning, "Stanford typed dependencies manual," Stanford University, Stanford, CA, USA, Tech. Rep., 2008.

[35] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *CoRR*, vol. abs/1301.3781, 2013.

[36] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of ACL 2007*, Prague, Czech Republic, 2007.

[37] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit." in *Proceedings of ICSLP 2002*, Denver, CO, USA, 2002.

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

[39] A. Venugopal, A. Zollman, and A. Waibel, "Training and Evaluation Error Minimization Rules for Statistical Machine Translation," in *Proceedings of WPT 2005*, Ann Arbor, MI, USA, 2005.

[40] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation." IBM Research Division, T. J. Watson Research Center, Tech. Rep. RC22176 (W0109-022), 2002.

[41] 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 IWSLT 2014*, Lake Tahoe, USA, December 2014.

[42] E. Cho, T.-L. Ha, J. Niehues, T. Herrmann, M. Mediani, Y. Zhang, and A. Waibel, "The Karlsruhe Institute of Technology Translation Systems for the WMT 2015," in *Proceedings of WMT 2015*, Lisbon, Portugal, September 2015.

[43] T.-L. Ha, J. Niehues, E. Cho, M. Mediani, and A. Waibel, "The KIT Translation Systems for IWSLT 2015," in *Proceedings of IWSLT 2015*, Da Nang, Vietnam, December 2015.

[44] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to Rank: From Pairwise Approach to Listwise Approach," in *Proceedings of ICML 2007*, Corvalis, OR, USA, 2007.