Stripping Adjectives: Integration Techniques for Selective Stemming in SMT Systems

Isabel Slawik **Jan Niehues** **Alex Waibel**
Institute for Anthropomatics and Robotics
KIT - Karlsruhe Institute of Technology, Germany
firstname.lastname@kit.edu

Abstract

In this paper we present an approach to reduce data sparsity problems when translating from morphologically rich languages into less inflected languages by selectively stemming certain word types. We develop and compare three different integration strategies: replacing words with their stemmed form, combined input using alternative lattice paths for the stemmed and surface forms and a novel hidden combination strategy, where we replace the stems in the stemmed phrase table by the observed surface forms in the test data. This allows us to apply advanced models trained on the surface forms of the words.

We evaluate our approach by stemming German adjectives in two German→English translation scenarios: a low-resource condition as well as a large-scale state-of-the-art translation system. We are able to improve between 0.2 and 0.4 BLEU points over our baseline and reduce the number of out-of-vocabulary words by up to 16.5%.

1 Introduction

Statistical machine translation (SMT) is currently the most promising approach to automatically translate text from one natural language into another. While it has been successfully used for a lot of languages and applications, many challenges still remain. Translating from a morphologically rich language is one such challenge where the translation quality of modern systems is often still not sufficient for many applications.

Traditional SMT approaches work on a lexical level, that is every surface form of a word is treated as its own distinct token. This can create data sparsity problems for morphologically rich languages, since the occurrences of a word are distributed over all its different surface forms. This problem becomes even more apparent when translating from an under-resourced language, where parallel training data is scarce.

When we translate from a highly inflected language into a less morphologically rich language, not all syntactic information encoded in the surface forms may be needed to produce an accurate translation. For example, verbs in French must agree with the noun in case and gender. When we translate these verbs into English, case and gender information may be safely discarded.

We therefore propose an approach to overcome these sparsity problems by stemming different morphological variants of a word prior to translation. This allows us to not only estimate translation probabilities more reliably, but also to translate previously unseen morphological variants of a word, thus leading to a better generalization of our models. To fully maximize the potential of our SMT system, we looked at three different integration strategies. We evaluated hard decision stemming, where all adjectives are replaced by their stem, as well as soft integration strategies, where we consider the words and their stemmed form as translation alternatives.

2 Related Work

The specific challenges arising from the translation of morphologically rich languages have been widely studied in the field of SMT. The factored translation model (Koehn and Hoang, 2007) enriches phrase-based MT with linguistic information. By translating the stem of a word and its morphological components separately and then applying generation rules to form the correct surface form of the target word, it is possible to generate translations for surface forms that have not been seen in training.

Talbot and Osborne (2006) address lexical redundancy by automatically clustering source words with similar translation distributions, whereas Yang and Kirchhoff (2006) propose a backoff model that uses increasing levels of morphological abstractions to translate previously unseen word forms.

Niehues and Waibel (2011) present quasi-morphological operations as a means to translate out-of-vocabulary (OOV) words. The automatically learned operations are able to split off potentially inflected suffixes, look up the translation for the base form using a lexicon of Wikipedia[1] titles in multiple languages, and then generate the appropriate surface form on the target side. Similar operations were learned for compound parts by Macherey et al. (2011).

Hardmeier et al. (2010) use morphological reduction in a German→English SMT system by adding the lemmas of every word output as a by-product of compound splitting as an alternative edge to input lattices. A similar approach is used by Dyer et al. (2008) and Wuebker and Ney (2012). They used word lattices to represent different source language alternatives for Arabic→English and German→English respectively.

Weller et al. (2013a) employ morphological simplification for their French→English WMT system, including replacing inflected adjective forms with their lemma using hand-written rules, and their Russian→English (Weller et al., 2013b) system, removing superfluous attributes from the highly inflected Russian surface forms. Their systems are unable to outperform the baseline system trained on the surface forms. Weller et al. argue that human translators may prefer the morphologically reduced system due to better generalization ability. Their analysis showed the Russian system often produces an incorrect verb tense, which indicates that some morphological information may be helpful to choose the right translation even if the information seems redundant.

3 Stemming

In order to address the sparsity problem, we try to cluster words that have the same translation probability distribution, leading to higher occurrence counts and therefore more reliable translation statistics. Because of the respective morphological properties of our source and target language, word stems pose a promising type of cluster. Moreover, stemming alleviates the OOV problem for unseen morphological variants. Because of these benefits, we chose stem clustering in this paper, however, our approach can work on different types of clusters, e.g. synonyms.

Morphological stemming prior to translation has to be done carefully, as we are actively discarding information. Indiscriminately stemming the whole source corpus hurts translation performance, since stemming algorithms make mistakes and often too much information is lost.

Adding the stem of every word as an alternative to our source sentence greatly increases our search space. Arguably the majority of the time we need the surface form of a word to make an informed translation decision. We therefore propose to keep the search space small by only stemming selected word classes which have a high diversity in inflections and whose additional morphological information content can be safely disregarded.

For our use case of translating from German to English, we chose to focus only on stemming adjectives. Adjectives in German can have five different suffixes, depending on the gender, number and case of the corresponding noun, whereas in English adjectives are only rarely inflected. We can therefore discard the information encoded in the suffix of a German adjective without losing any vital information for translation.

3.1 Degrees of Comparison

While we want to remove gender, number and case information from the German adjective, we want to preserve its comparative or superlative nature. In addition to its base form (e.g. *schön* [*pretty*]), a German adjective can have one of five suffixes (*-e, -em, -en, -er, -es*). However, we cannot simply remove all suffixes using fixed rules, because the comparative base form of an adjective is identical to the inflected masculine, nominative, singular form of an attributive adjective.

For example, the inflected form *schöner* of the adjective *schön* is used as an attributive adjective in

[1] http://www.wikipedia.org

the phrase *schöner Mann* [*handsome man*] and as a comparative in the phrase *schöner wird es nicht* [*won't get prettier*]. We can stem the adjective in the attributive case to its base form without any confusion (*schön Mann*), as we generate a form that does not exist in proper German. However, were we to apply the same stemming to the comparative case, we would lose the degree of comparison and still generate a valid German sentence (*schön wird es nicht* [*won't be pretty*]) with a different meaning than our original sentence. In order to differentiate between cases in which stemming is desirable and where we would lose information, a detailed morphological analysis of the source text prior to stemming is vital.

3.2 Implementation

We used readily available part-of-speech (POS) taggers, namely the TreeTagger (Schmid, 1994) and RFTagger (Schmid and Laws, 2008), for morphological analysis and stemming. In order to achieve accurate results, we performed standard machine translation preprocessing on our corpora before tagging. We discarded exceedingly long sentences and sentence pairs with a large length difference from the training data. Special dates, numbers and symbols were normalized and we smart-cased the first letter of every sentence. Typically preprocessing for German also includes splitting up compounds into their separate parts. However, this would confuse the POS taggers, which have been trained on German text with proper compounds. Furthermore, our compound splitting algorithm might benefit from a stemmed corpus, providing higher occurrence counts for individual word components. We therefore refrain from compound splitting before tagging and stemming.

We only stemmed words tagged as attributive adjectives, since only they are inflected in German. Predicative adjectives are not inflected and therefore were left untouched. Since we want to retain the degree of comparison, we used the fine-grained tags of the RFTagger to decide when and how to stem. Adjectives tagged as comparative or superlative were stemmed through the use of fixed rules. For all others, we used the lemma output by the TreeTagger, since it is the same as the stem and was already available in our system.

Finally, our usual compound splitting (Koehn and Knight, 2003) was trained and performed on the stemmed corpus.

4 Integration

After clustering the words into groups that can be translated in the same or at least in a similar way, there are different possibilities to use them in the translation system. A naive strategy is to replace each word by its cluster representative, called *hard decision stemming*. However, this carries the risk of discarding vital information. Therefore we investigated techniques to integrate both, the surface forms as well as the word stems, into the translation system. In the *combined input*, we add the stemmed adjectives as translation alternatives to the preordering lattices. Since this poses problems for the application of more advanced translation models during decoding, we propose the novel *hidden combination* technique.

4.1 Hard Decision Stemming

Assuming that the translation probabilities of the word stems can be estimated more reliably than those of the surface forms, the most intuitive strategy is to consequently replace each surface form by its stem. In our case, we replaced all adjectives with their stems. This has the advantage that afterwards the whole training pipeline can be performed in exactly the same manner as it is done in the baseline system. For tuning and testing, the adjectives in the development and test data are stemmed and replaced in the same manner as in the training data.

4.2 Combined Input

Mistakes made during hard decision stemming cannot be recovered. Soft integration techniques avoid this pitfall by deferring the decision whether to use the stem or surface form of a word until decoding. We enable our system to choose by combining both the surface form based (default) phrase table and the word stem based (stemmed) phrase table log-linearly. The weights of the phrase scores are then learned during optimization.

In order to be able to apply both phrase tables at the same time, we need to modify the input of the decoder. Our baseline system already uses preordering lattices, which encode different reordering possibilities of the source sentence. We replaced every edge in the lattice containing an adjective by two edges: one containing the surface form and the other the word stem. This allows the decoder to choose which word form to use depending on the word and its context.

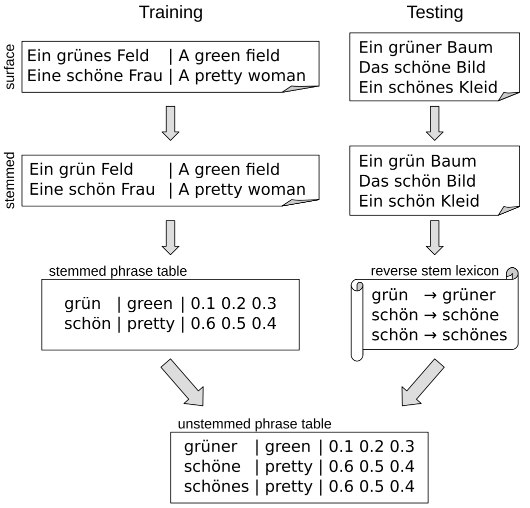
Training
Testing
surface
Ein grünes Feld | A green field
Eine schöne Frau | A pretty woman
Ein grüner Baum
Das schöne Bild
Ein schönes Kleid
stemmed
Ein grün Feld | A green field
Eine schön Frau | A pretty woman
Ein grün Baum
Das schön Bild
Ein schön Kleid
stemmed phrase table
grün | green | 0.1 0.2 0.3
schön | pretty | 0.6 0.5 0.4
reverse stem lexicon
grün → grüner
schön → schöne
schön → schönes
unstemmed phrase table
grüner | green | 0.1 0.2 0.3
schöne | pretty | 0.6 0.5 0.4
schönes | pretty | 0.6 0.5 0.4

Figure 1: Workflow for unstemming the PT.

4.3 Hidden Combination

While we are able to modify our phrase table to use both surface forms and stems in the last strategy, other models in our log-linear system suffer from the different types of source input. For example, the bilingual language model (Niehues et al., 2011) is based on tokens of target words and their aligned source words. In training, we can use either the stemmed corpus or the original one, but during decoding a mixture of stems and surface forms occurs. For the unknown word forms the scores will not be accurate and the performance of our model will suffer. Similar problems occur when using other translation models such as neural network based translation models.

We therefore developed a novel strategy to integrate the word stems into the translation system. Instead of stemming the input to fit the stemmed phrase table, we modified the stemmed phrase table so that it can be applied to the surface forms. The workflow is illustrated in Figure 1. We extracted all the stem mappings from the development and test data and compiled a stem lexicon. This maps the surface forms observed in the dev and test data to their corresponding stems. We then applied this lexicon in reverse to our stemmed phrase table, in effect duplicating every entry containing a stemmed adjective with the inflected form replacing the stem. Afterwards this "unstemmed" phrase table is log-linearly combined with the default phrase table and used for translation.

This allows us to retain our generalization won by using word clusters to estimate phrase probabilities, and still use all models trained on the surface forms. Using the hidden combination strategy, stemming can easily be implemented into current state-of-the-art SMT systems without the need to change any of the advanced models beyond the phrase table. This makes our approach highly versatile and easy to implement for any number of system architectures and languages.

5 Experiments

Since we expect stemming to have a larger impact in cases where training data is scarce, we evaluated the three presented strategies on two different scenarios: a low-resource condition and a state-of-the-art large-scale system. In both scenarios we stemmed German adjectives and translated from German to English.

In our low-resource condition, we trained an SMT system using only training data from the TED corpus (Cettolo et al., 2012). TED translations are currently available for 107 languages[2] and are being continuously expanded. Therefore, there is a high chance that a small parallel corpus of translated TED talks will be available in the chosen language.

In the second scenario, we used a large-scale state-of-the-art German→English translation system. This system was trained on significantly more data than available in the low-resource condition and incorporates several additional models.

5.1 System Description

The low-resource system was trained only on the TED corpus provided by the IWSLT 2014 machine translation campaign, consisting of 172k lines. As monolingual training data we used the target side of the TED corpus.

The large-scale system was trained on the European Parliament Proceedings, News Commentary, TED and Common Crawl corpora provided for the IWSLT 2014 machine translation campaign (Cettolo et al., 2014), encompassing 4.69M lines. For the monolingual training data we used the target side of all bilingual corpora as well as the News Shuffle and the Gigaword corpus.

Before training and translation, the data is preprocessed as described in Section 3.2. The noisy Common Crawl corpus was filtered with an SVM classifier as described by Mediani et al. (2011). After preprocessing, the parallel corpora are word-aligned with the GIZA++ toolkit (Gao and Vo-

[2] http://www.ted.com/participate/translate

gel, 2008) in both directions. The resulting alignments are combined using the *grow-diag-final-and* heuristic. The Moses toolkit (Koehn et al., 2007) is used for phrase extraction. For the large-scale system, phrase table adaptation combining an in-domain and out-of-domain phrase table is performed (Niehues and Waibel, 2012). All translations are generated by our in-house phrase-based decoder (Vogel, 2003).

We used 4-gram language models (LMs) with modified Kneser-Ney smoothing, trained with the SRILM toolkit (Stolcke, 2002) and scored in the decoding process with KenLM (Heafield, 2011).

All our systems include a reordering model which automatically learns reordering rules based on part-of-speech sequences and, in case of the large-scale system, syntactic parse tree constituents to better match the target language word order (Rottmann and Vogel, 2007; Niehues and Kolss, 2009; Herrmann et al., 2013). The resulting reordering possibilities for each source sentence are encoded in a lattice.

For the low-resource scenario, we built two systems. One small baseline with only one phrase table and language model, as well as aforementioned POS-based preordering model, and an advanced system using an extended feature set of models that are also used in the large-scale system. The extended low-resource and the large-scale system include the following additional models.

A bilingual LM (Niehues et al., 2011) is used to increase the bilingual context during translation beyond phrase boundaries. It is built on tokens consisting of a target word and all its aligned source words. We also used a 9-gram cluster LM built on 100 automatically clustered word classes using the MKCLS algorithm (Och, 1999).

The large-scale system also uses an in-domain LM trained on the TED corpus and a word-based model trained on 10M sentences chosen through data selection (Moore and Lewis, 2010).

In addition to the lattice preordering, a lexicalized reordering model (Koehn et al., 2005) which stores reordering probabilities for each phrase pair is included in both extended systems.

We tune all our systems using MERT (Venugopal et al., 2005) against the BLEU score. Since the systems have a varying amount of features, we reoptimized the weights for every experiment.

For the low-resource system, we used IWSLT test 2012 as a development set and IWSLT test 2011 as test data. For the large-scale system, we used IWSLT test 2011 as development data and IWSLT test 2012 as test data.

All results are reported as case-sensitive BLEU scores calculated with one reference translation.

System	Dev	Test
Baseline	28.91	30.25
Hard Decision	29.01	30.30
Combined Input	29.13	30.47
Hidden Combination	**29.25**	**30.62**

Table 1: TED low-resource small systems results.

5.2 Low-resource Condition

The results for the systems built only on the TED corpus are summarized in Table 1 for the small system and Table 2 for the extended system. The baseline systems reach a BLEU score on the test set of 30.25 and 31.33 respectively.

In the small system we could slightly improve to 30.30 using only stemmed adjectives. However, in the extended system the hard decision strategy could not outperform the baseline. This indicates that for words with sufficient data it might be better to translate the surface forms.

Adding the stemmed forms as alternatives to the preordering lattice leads to an improvement of 0.2 BLEU points over the small baseline system. In the larger system with the extended features set, the combined input performed better than the hard decision stemming, but is still 0.1 BLEU points below the baseline. With this strategy we do not tap the full potential of our extended system, as there is still a mismatch between the combined input and the training data of the advanced models.

The hidden combination strategy rectifies this problem, which is reflected in the results. Using the hidden combination we could achieve our best BLEU score for both systems. We could improve by almost 0.4 BLEU points over the small baseline system and 0.3 BLEU points on the system using extended features.

System	Dev	Test
Baseline	29.73	31.33
Hard Decision	29.74	30.84
Combined Input	**29.97**	31.22
Hidden Combination	29.87	**31.61**

Table 2: TED extended features systems results.

System	Dev	Test
Baseline	38.30	30.89
Hard Decision	38.25	30.82
Combined Input	**38.65**	**31.10**
Hidden Combination	38.40	31.08

Table 3: IWSLT large-scale systems results.

5.3 Large-scale System

In order to assess the impact of our stemming on a state-of-the-art system, we tested our techniques on a large-scale system using training data from several domains. The results of these experiments are summarized in Table 3. The baseline system achieved a BLEU score of 30.89 on the test set.

As in the low-resource condition, the hard decision to use only the stems causes a slight drop in performance. Given the large amount of training data, the problem of having seen a word few times is much less severe than before.

When we combine the inputs, we can improve the translation quality to our best score of 31.10 BLEU points. The hidden combination performs similarly. By using combined input or hidden combination, we achieved a gain of 0.2 BLEU points over the baseline.

5.4 Further Analysis

In this work we have focused on selectively stemming only a small subset of our input text, namely adjectives. We therefore do not expect to see a large difference in BLEU score in our systems and indeed the improvements, while existent, are moderate. It is a well known shortcoming of automatic metrics that they cannot differentiate between acceptable translation alternatives and errors. Since time and monetary constraints did not allow us to perform a full-scale human evaluation, we use the OOV rate and manual inspection to demonstrate the benefits of our approach.

For a monolingual user of machine translation systems, even an imperfect translation will be better than no translation at all. We therefore looked at the out-of-vocabulary (OOV) rate of our systems.

477 OOV words occurred in the test set of the low-resource baseline. This means of the 1433 lines in our test set, on average every third contained an untranslated word. With stemming we were able to translate 79 of those words and reduce the number of OOV words by 16.5%. Even in the large-scale system, which is trained on a large amount of data and therefore has an already low OOV rate, we achieved a decrease of 4%. Figure 2 shows an example sentence where we managed to translate two previously OOV words using the hidden combination strategy. Furthermore, stemming can also improve our word choices as shown in the example in Figure 3.

SRC	Aber es war sehr traurig .
REF	But it was very sad .
BASE	But it was really upset .
H.C.	But it was very sad .

Figure 3: Example of improved word choice.

Stemming certain words in a corpus not only affects the translation of that word, but the whole system. For example, stemming changes the occurrence statistics of the stemmed words, and therefore the output of empirical algorithms such as compound splitting and word alignment is subject to change. By combining the stemmed and default phrase tables, we gave our decoder the chance to use a phrase from the stemmed phrase table even if the phrase contains no stemmed words. A manual evaluation of the output of the hidden combination system compared to the hard decision stemmed system showed that the difference was largely in word order as exemplified in Figure 4.

6 Conclusion

In this paper we addressed the problem of translating from morphologically rich languages into less inflected languages. The problem of low occur-

SRC	Während Schimpansen von großen , **furchteinflößenden** Kerlen geführt werden , wird die Bonobo - Gesellschaft von **ermächtigten** Weibchen geführt .
REF	While chimpanzees are dominated by big , **scary** guys , bonobo society is run by **empowered** females .
BASE	As chimpanzees by large , **fear einflößenden** guys are , the Bonobo-society led by **ermächtigten** females .
H.C.	During the chimpanzees of big , **scary** guys are , the Bonobo is society of **empowered** females .

Figure 2: Example translations of the baseline and hidden combination low-resource systems. OOV phrases have been marked in bold.

SRC	Nun ja , eine Erleuchtung ist für gewöhnlich etwas , dass man findet weil man es irgendwo fallen gelassen hat .
REF	And you know , an epiphany is usually something you find that you dropped someplace .
H.D.	Well , there is an epiphany usually , something that you can find because it has somewhere dropped .
H.C.	Well , an epiphany is usually something that you can find because it has dropped somewhere .

Figure 4: Example of improved word order of the hidden combination over the hard decision system.

rence counts for surface forms and high out-of-vocabulary rates for unobserved surface forms can be alleviated by stemming words.

We showed that stemming has to be done carefully, since SMT systems are highly sensitive to lost information. Given our use case of German to English translation, we chose to only stem adjectives, which can have five suffixes depending on gender, number and case of the corresponding noun. We took special care to ensure comparative and superlative adjectives retained their degree of comparison after stemming.

As an alternative to the hard decision strategy, where every word is replaced by its stem, we proposed two soft integration techniques incorporating the stems and surface forms as alternative translation paths in the preordering lattices. State-of-the-art SMT systems consist of a log-linear combination of many advanced models. Combining the surface forms and word stems posed problems for models relying on source side tokens. We therefore developed a novel hidden combination technique, where the word stems in the phrase table are replaced by the observed surface forms in the test data. This allowed us to use the more reliably estimated translation probabilities calculated on the word stems in the decoder while simultaneously applying all our other models to the surface forms of the words.

We evaluated our approach on German→English translation in two scenarios, one low-resource condition and a large-scale state-of-the-art SMT system. Given the low-resource condition, we evaluated a small, basic system as well as a more sophisticated system using an extended feature set. Using the hidden combination strategy, we were able to outperform the baseline systems in all three experiments by 0.2 up to 0.4 BLEU points. While these improvements may seem moderate, they were achieved solely through the modification of adjectives. We were also able to show that our systems generalized better than the baseline as evidenced by the OOV rate, which could be decreased by 16.5% in the low-resource condition.

Acknowledgments

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.

References

Cettolo, M., C. Girardi, and M. Federico. 2012. WIT3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Annual Meeting of the European Association for Machine Translation*, Trento, Italy.

Cettolo, M., J. Niehues, S. Stüker, L. Bentivogli, and M. Federico. 2014. Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014. In *Proceedings of the 11th International Workshop on Spoken Language Translation*, Lake Tahoe, California, USA.

Dyer, C., S. Muresan, and P. Resnik. 2008. Generalizing Word Lattice Translation. In *Proceedings of the 46th Annual Meeting of the ACL: Human Language Technologies*, Columbus, Ohio, USA.

Gao, Q. and S. Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Proceedings of the Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, Columbus, Ohio, USA.

Hardmeier, C., A. Bisazza, M. Federico, and F.B. Kessler. 2010. FBK at WMT 2010: Word Lattices for Morphological Reduction and Chunk-based Reordering. In *Proceedings of the Fifth Workshop on Statistical Machine Translation and Metrics MATR*, Uppsala, Sweden.

Heafield, K. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, UK.

Herrmann, T., J. Niehues, and A. Waibel. 2013. Combining Word Reordering Methods on Different Linguistic Abstraction Levels for Statistical Machine Translation. In *Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, Atlanta, Georgia, USA.

Koehn, P. and H. Hoang. 2007. Factored Translation Models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic.

Koehn, P. and K. Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of the 10th Conference of the European Chapter of the ACL*, Budapest, Hungary.

Koehn, P., A. Axelrod, A.B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *Proceedings of the 2nd International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Koehn, P., 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. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL*, Prague, Czech Republic.

Macherey, K., A. Dai, D. Talbot, A. Popat, and F. Och. 2011. Language-independent Compound Splitting with Morphological Operations. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, Portland, Oregon, USA.

Mediani, M., E. Cho, J. Niehues, T. Herrmann, and A. Waibel. 2011. The KIT English-French Translation Systems for IWSLT 2011. In *Proceedings of the Eights International Workshop on Spoken Language Translation*, San Francisco, California, USA.

Moore, R.C. and W. Lewis. 2010. Intelligent Selection of Language Model Training Data. In *Proceedings of the 48th Annual Meeting of the ACL*, Uppsala, Sweden.

Niehues, J. and M. Kolss. 2009. A POS-based Model for Long-Range Reorderings in SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece.

Niehues, J. and A. Waibel. 2011. Using Wikipedia to Translate Domain-Specific Terms in SMT. In *Proceedings of the Eights International Workshop on Spoken Language Translation*, San Francisco, California, USA.

Niehues, J. and A. Waibel. 2012. 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*, San Diego, California, USA.

Niehues, J., T. Herrmann, S. Vogel, and A. Waibel. 2011. Wider Context by Using Bilingual Language Models in Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, UK.

Och, F.J. 1999. An Efficient Method for Determining Bilingual Word Classes. In *Proceedings of the Ninth Conference of the European Chapter of the ACL*, Bergen, Norway.

Rottmann, K. and S. Vogel. 2007. 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.

Schmid, H. and F. Laws. 2008. Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, UK.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Stolcke, A. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference of Spoken Language Processing*, Denver, Colorado, USA.

Talbot, D. and M. Osborne. 2006. Modelling Lexical Redundancy for Machine Translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, Sydney, Australia.

Venugopal, A., A. Zollman, and A. Waibel. 2005. Training and Evaluation Error Minimization Rules for Statistical Machine Translation. In *Proceedings of the Workshop on Data-driven Machine Translation and Beyond*, Ann Arbor, Michigan, USA.

Vogel, S. 2003. SMT Decoder Dissected: Word Reordering. In *Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China.

Weller, M., A. Fraser, and S. Schulte im Walde. 2013a. Using Subcategorization Knowledge to Improve Case Prediction for Translation to German. In *Proceedings of the 51st Annual Meeting of the ACL*, Sofia, Bulgaria.

Weller, M., M. Kisselew, S. Smekalova, A. Fraser, H. Schmid, N. Durrani, H. Sajjad, and R. Farkas. 2013b. Munich-Edinburgh-Stuttgart Submissions of OSM Systems at WMT13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria.

Wuebker, J. and H. Ney. 2012. Phrase Model Training for Statistical Machine Translation with Word Lattices of Preprocessing Alternatives. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada.

Yang, M. and K. Kirchhoff. 2006. Phrase-Based Backoff Models for Machine Translation of Highly Inflected Languages. In *Proceedings of the 11th Conference of the European Chapter of the ACL*, Trento, Italy.