NECTAR: Modeling Inflection and Word-Formation in SMT

Alexander Fraser[1], Marion Weller[2], Aoife Cahill[3], Fabienne Cap[2]

[1] CIS, Ludwig-Maximilians-Universität München
[2] IMS, Universität Stuttgart
[3] Educational Testing Service, Princeton, USA

Abstract. The current state-of-the-art in statistical machine translation (SMT) suffers from issues of sparsity and inadequate modeling power when translating into morphologically rich languages. We model both inflection and word-formation for the task of translating into German. We translate from English words to an underspecified German representation and then use linear-chain CRFs to predict the fully specified German representation. We show that improved modeling of inflection and word-formation leads to improved SMT.

Keywords: statistical machine translation, inflection, word-formation

Summary: This is a NECTAR contribution summarizing and discussing [6]. We model both inflection and word-formation for the task of translating into German using statistical machine translation. We translate from English words to an underspecified German representation and then use linear-chain CRFs to predict the fully specified German representation. We show that improved modeling of inflection and word-formation leads to improvements in translation.

Phrase-based statistical machine translation suffers from problems of data sparsity with respect to inflection and word-formation which are particularly strong when translating to a morphologically rich target language, such as German. Our work is important as it combines two research directions in the SMT literature which are usually examined independently of one another: generation of inflection and dealing with problems of word-formation. Our system addresses both of these issues, which are clearly related to one another.

We address the problem of inflection by first translating to a stem-based representation, and then using a second process to inflect these stems. We study several models for doing this including: strongly lexicalized models, unlexicalized models using linguistic features, and models combining the strengths of both of these approaches. We address the problem of word-formation for compounds in German, by translating from English into German word parts, and then determining whether to merge these parts to form compounds. We make the following new contributions: i) We introduce the first SMT system combining inflection prediction with synthesis of compounds (we also handle synthesis of portmanteaus, but more importantly we deal with issues of inflection involving portmanteaus in a consistent way with other issues of inflection). ii) For

inflection, we compare the mostly unlexicalized prediction of linguistic features (with a subsequent surface form generation step) versus the direct prediction of surface forms, and show that both approaches have complementary strengths. iii) We combine the advantages of the prediction of linguistic features with the prediction of surface forms. We implement this in a CRF framework which improves on a standard phrase-based SMT baseline. iv) We develop separate (but related) procedures for inflection prediction and dealing with word-formation (compounds and portmanteaus), in contrast with most previous work which usually either tried to solve both problems using approaches appropriate for inflectional problems, or tried to solve both problems using approaches appropriate for word-formation problems.

We implement: i) generalization of nominal inflection (for instance, the English NP "the respective countries" is translated into German using three different words depending on whether the case is nominative or dative, see Table 1). ii) generalization over highly productive noun compounding, such as "developing countries", which is usually a single word in German. iii) generalization over portmanteaus consisting of a preposition plus an article (prevalent in German and dependent on the linguistic feature *case*), to be able to correctly translate a noun only seen together with a portmanteau in the training data to a non-portmanteau construction and vice versa.

By combining the use of the BitPar parser [13] with the use of the Stuttgart Morphological Analyzer SMOR [14], we were able to obtain high quality analyses of the German text in our parallel data and in the larger monolingual data used for German language modeling. We designed two lexicon representations for German (and tested using them for inflection prediction).

The first lexicon representation involves splitting compounds using the techniques developed by [7], and using morphological disambiguation from BitPar to pick the correct SMOR entries for all inflected words (thereby reducing them to lemma + morphological features). The second lexicon representation is an enhanced version of this and will be described below.

After designing the first representation, we performed studies on a two-step translation process. We first translate from English text to an underspecified representation, and then in the second step we create the fully specified representation, which maps directly to a German string realization. We focused on nominal morphology. We translated English to a stem-like representation of German articles, adjectives and nouns. We initially used an HMM-like prediction process which took stems as input and produced inflected words as output. We then switched to prediction of the four linguistic features *case*, *gender*, *number*, and *in-weak-context* (strong/weak adjectival inflection), and designed a system using SMOR to generate the final inflected forms.

After further experimentation we realized that using only word stems was not a sufficient representation as input to the inflection prediction. We therefore designed a stem markup, where we annotate stems with some linguistic features (so, for instance, German prepositions are marked with the *case* they take, e.g., "auf+Dat" means a usage of the preposition "auf" with a Dative object). This

output decoder	input prediction	output prediction	inflected forms	gloss
haben<VAFIN>	haben-V	haben-V	haben	*have*
Zugang<+NN><Masc><Sg>	NN-Sg-Masc	NN-Masc.Acc.Sg.iwc=0	Zugang	*access*
zu<APPR><Dat>	APPR-zu-Dat	APPR-zu-Dat	zu	*to*
die<+ART><Def>	ART-def	ART-Neut.Dat.Pl.iwc=1	den	*the*
betreffend<+ADJ><Pos>	ADJA	ADJA-Neut.Dat.Pl.iwc=1	betreffenden	*respective*
Land<+NN><Neut><Pl>	NN-Pl-Neut	NN-Neut.Dat.Pl.iwc=1	Ländern	*countries*

Table 1. Overview: inflection prediction steps using a single joint sequence model. All words except verbs and prepositions are replaced by their POS tags in the input. Verbs are inflected in the input ("haben", meaning "have" as in "they have", in the example). Prepositions are lexicalized ("zu" in the example) and indicate which *case* value they mark ("Dat", i.e., Dative, in the example). The annotation iwc=0 indicates "in-weak-context" is false.

stem markup is the second lexicon representation. Given this stem markup representation as input, the inflection prediction process is much more effective. We determined that it was not much more difficult for a phrase-based SMT decoder to predict stems+markup than it was to predict only the stems (yet this was far easier than predicing the correct surface form), validating our choice. Further examples of the stem markup are shown in the input column of Table 1. For additional discussion of the stem markup, see [5].

An overview of the prediction process for a single joint model of all four features is shown in Table 1. The best performing pipeline in [6] is using four linear-chain CRFs to predict each linguistic feature separately. This works because the linguistic features marked in the stemmed input to inflection prediction have enough information to enable this independence.

We evaluate on the end-to-end SMT task of translating from English to German of the 2009 ACL workshop on SMT. We achieve statistically significant BLEU score increases on both the test set and the blind test set. We also implemented translation to a split compound representation, and used a CRF to make the binary decision of where to merge words to form compounds, see [6].

Discussion: The idea of translating to stems and then inflecting is not novel. We adapted [17] which is effective but limited by the conflation of two separate issues: word-formation and inflection. Given a stem such as *brother*, such a system might generate the "stem and inflection" corresponding to *and his brother*. Viewing *and* and *his* as inflection is problematic since a mapping from the English phrase *and his brother* to the Arabic stem for *brother* is required. The situation is worse if there are English words (e.g., adjectives) separating *his* and *brother*. This required mapping is a significant problem for generalization. We view this issue as a different sort of problem entirely, one of word-formation (rather than inflection). We apply a "split in preprocessing and resynthesize in postprocessing" approach to these phenomena, combined with inflection prediction that is similar to that of [17]. The only work that we are aware of which deals with both issues is [8] which deals with verbal morphology and attached pronouns. There has been other work on solving inflection. [9] introduced factored SMT. We use more complex context features. [4] tried to solve the inflection predic-

tion problem by simply building an SMT system for translating from stems to inflected forms. [2] improved on this by marking prepositions with the *case* they mark (one of the most important markups in our system). Both efforts were ineffective on large data sets. [20] used unification in an SMT system to model some of the agreement phenomena that we model. Our CRF framework allows us to use more complex context features.

We have directly addressed the question as to whether inflection should be predicted using surface forms as the target of the prediction, or whether linguistic features should be predicted, along with the use of a subsequent generation step. The direct prediction of surface forms is limited to those forms observed in the training data, which is a significant limitation. However, it is reasonable to expect that the use of features (and morphological generation) could also be problematic as this requires the use of morphologically-aware syntactic parsers to annotate the training data with such features, and additionally depends on the coverage of morphological analysis and generation. Despite this, our research clearly shows that the feature-based approach is superior for English-to-German SMT. This is a striking result considering state-of-the-art performance of German parsing is poor compared with the best performance on English parsing. As parsing performance improves, linguistic-feature-based approaches will perform better. We found the prediction of *case* to be the most difficult, see [19] for recent work on modeling verb subcategorization for this task.

[18], [1], [11], [3], and others are primarily concerned with using morpheme segmentation in SMT, which is a useful approach for dealing with issues of word-formation. However, this does not deal directly with linguistic features marked by inflection. In German these linguistic features are marked very irregularly and there is widespread syncretism, making it difficult to split off morphemes specifying these features. So it is questionable as to whether morpheme segmentation techniques are sufficient to solve the inflectional problem we are addressing.

For compound splitting, we follow [7], using linguistic knowledge encoded in a rule-based morphological analyser and then selecting the best analysis based on the geometric mean of word part frequencies. Other approaches use less deep linguistic resources (e.g., POS-tags [15]) or are (almost) knowledge-free (e.g., [10]). Compound merging is less well studied. [12] used a simple, list-based merging approach, merging all consecutive words included in a merging list. This approach resulted in too many compounds. We follow [16], for compound merging. We trained a CRF using (nearly all) of the features they used and found their approach to be effective (when combined with inflection and portmanteau merging) on one of our two test sets.

Conclusion: We have shown that both the prediction of surface forms and the prediction of linguistic features are of interest for improving SMT. We have obtained the advantages of both in our CRF framework, and also integrated handling of compounds, and an inflection-dependent word-formation phenomenon, portmanteaus. We validated our work on a well-studied large corpora translation task. In future work, we plan to improve our compound merging system further and expand our system to handle verbal inflection.

Acknowledgments

The authors wish to thank the anonymous reviewers for their comments. Aoife Cahill was partly supported by Deutsche Forschungsgemeinschaft grant SFB 732. Alexander Fraser, Marion Weller and Fabienne Cap were funded by Deutsche Forschungsgemeinschaft grant Models of Morphosyntax for Statistical Machine Translation. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement Nr. 248005. This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

References

1. Badr, I., Zbib, R., Glass, J.: Segmentation for English-to-Arabic statistical machine translation. In: Proceedings of ACL-08: HLT, Short Papers. pp. 153–156 (2008)
2. Bojar, O., Kos, K.: 2010 Failures in English-Czech Phrase-Based MT. In: Proc. of WMT10/MetricsMATR (2010)
3. Clifton, A., Sarkar, A.: Combining morpheme-based machine translation with post-processing morpheme prediction. In: ACL. pp. 32–42 (2011)
4. Fraser, A.: Experiments in morphosyntactic processing for translating to and from German. In: EACL WMT. pp. 115–119 (2009)
5. Fraser, A., Weller, M., Cahill, A., Cap, F.: Morphological generation of German for SMT. In: Machine Translation and Morphologically-rich Languages. Workshop of Israel Science Foundation (2011)
6. Fraser, A., Weller, M., Cahill, A., Cap, F.: Modeling inflection and word-formation in SMT. In: EACL. pp. 664–674 (2012)
7. Fritzinger, F., Fraser, A.: How to avoid burning ducks: Combining linguistic analysis and corpus statistics for German compound processing. In: ACL WMT and Metrics MATR. pp. 224–234 (2010)
8. de Gispert, A., Mariño, J.B.: On the impact of morphology in English to Spanish statistical MT. Speech Communication 50(11-12), 1034–1046 (2008)
9. Koehn, P., Hoang, H.: Factored translation models. In: EMNLP-CONLL (2007)
10. Koehn, P., Knight, K.: Empirical methods for compound splitting. In: EACL. pp. 187–193 (2003)
11. Luong, M.T., Nakov, P., Kan, M.Y.: A hybrid morpheme-word representation for machine translation of morphologically rich languages. In: EMNLP (2010)
12. Popović, M., Stein, D., Ney, H.: Statistical machine translation of German compound words. In: FinTAL - 5th International Conference on Natural Language Processing, Springer Verlag, LNCS. pp. 616–624 (2006)
13. Schmid, H.: Efficient parsing of highly ambiguous context-free grammars with bit vectors. In: COLING (2004)
14. Schmid, H., Fitschen, A., Heid, U.: SMOR: a German computational morphology covering derivation, composition, and inflection. In: LREC. pp. 1263–1266 (2004)
15. Stymne, S.: German compounds in factored statistical machine translation. In: GoTAL. pp. 464–475 (2008)
16. Stymne, S., Cancedda, N.: Productive Generation of Compound Words in Statistical Machine Translation. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 250–260 (2011)

17. Toutanova, K., Suzuki, H., Ruopp, A.: Applying morphology generation models to machine translation. In: ACL-HLT (2008)
18. Virpioja, S., Väyrynen, J.J., Creutz, M., Sadeniemi, M.: Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In: PROC. OF MT SUMMIT XI. pp. 491–498 (2007)
19. Weller, M., Fraser, A., Schulte im Walde, S.: Using subcategorization knowledge to improve case prediction for translation to German. In: ACL (2013)
20. Williams, P., Koehn, P.: Agreement constraints for statistical machine translation into german. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 217–226 (2011)