Predicting Prepositions for SMT

Marion Weller[1,2], **Alexander Fraser**[2], **Sabine Schulte im Walde**[1]

[1] IMS, University of Stuttgart – (wellermn|schulte)@ims.uni-stuttgart.de
[2] CIS, Ludwig-Maximilian University of Munich – fraser@cis.uni-muenchen.de

Introduction The translation of prepositions is a difficult task for machine translation; a preposition must convey the source-side meaning and also meet target-side constraints. In our approach, we move the selection of prepositions out of the translation system into a post-processing component. During translation, we use an abstract representation of prepositions as a place-holder that serves as a basis for the generation of prepositions in the post-processing step: all subcategorized elements of a verb are considered and allotted to their respective functions – as PPs with an overt preposition or as NPs with an "empty" preposition, e.g. *to call for sth.* → *∅ etw. erfordern*. The language model and the translation rules often fail to correctly model subcategorization in standard SMT systems because verbs and their subcategorized elements are often not adjacent.

We use a morphology-aware SMT system which first translates into a lemmatized representation with a component to generate fully inflected forms in a second step, see Toutanova et al. (2008) and Fraser et al. (2012). The inflection step requires the modeling of the grammatical *case* of noun phrases, which corresponds to determining the syntactic function. Weller et al. (2013) describe modeling *case* in SMT; we extend their setup to cover the prediction of prepositions in both PP and NPs (i.e., the "empty" preposition). The presented work is similar to that of Agirre et al. (2009), but is applied to a fully statistical MT system. A detailed presentation of our work including a full literature survey can be found in Weller et al. (2015).

Methodology To build the translation model, we use an abstract target-language representation in which nouns, adjectives and articles are lemmatized, and prepositions are substituted with place-holders. Additionally, "empty" place-holder prepositions are inserted at the beginning of noun phrases. To obtain a symmetric data structure, "empty" place-holders are also added to source-side NPs. When generating surface forms for the translation output, a phrase with a place-holder preposition can be realized as a noun phrase (empty preposition) or as a prepositional phrase by generating the preposition's surface form.

Figure 1 illustrates the process: for the English input with the extra null-prepositions (column 1), the SMT system outputs a lemmatized representation with place-holder prepositions (column 2). In a first step, prepositions and *case* for the SMT output are predicted (column 3). Then, the three remaining inflection-relevant morphological features *number*, *gender* and *strong/weak* are predicted on "regular" sentences without place-holders, given the prepositions from the previous step (column 4). In the last step, fully inflected forms are produced based on features and lemmas (column 5).

Abstract Representation and Prediction Features Initial experiments showed that replacing prepositions by simple place-holders decreases the translation quality. As an extension to the basic approach with plain place-holders, we thus experiment with enriching the place-holders such that they contain more relevant information and represent the content of a preposition while still being in an abstract form. For example, the representation can be enriched by annotating the place-holder with the grammatical case of the preposition it represents: for overt prepositions, case is often an indicator of the content (e.g. direction/location), whereas for NPs, case indicates

Proceedings of SSST-9, Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 55–56,
Denver, Colorado, June 4, 2015.

input	lemmatized SMT output	prep	morph. feat.	inflected	gloss
∅ ⟶	PREP	∅-Acc	–		
what	welch<PWAT>	Acc	Acc.Fem.Sg.Wk	welche	which
role	Rolle<+NN><Fem><Sg>	Acc	Acc.Fem.Sg.Wk	Rolle	role
∅ ⟶	PREP	∅-Nom	–		
the	die<+ART><Def>	Nom	Nom.Masc.Sg.St	der	the
giant	riesig<ADJ>	Nom	Nom.Masc.Sg.Wk	riesige	giant
planet	Planet<+NN><Masc><Sg>	Nom	Nom.Masc.Sg.Wk	Planet	planet
has	gespielt<VVPP>	–	–	gespielt	played
played	hat<VAFIN>	–	–	hat	has
in ⟶	PREP	bei-Dat	–	bei	for
the	die<+ART><Def>	Dat	Dat.Fem.Sg.St	der	the
development	Entwicklung<+NN><Fem><Sg>	Dat	Dat.Fem.Sg.Wk	Entwicklung	development
of ⟶	PREP	∅-Gen	–		
the	die<+ART><Def>	Gen	Gen.Neut.Sg.St	des	of-the
solar system	Sonnensystem<+NN><Neut><Sg>	Gen	Gen.Neut.Sg.Wk	Sonnensystems	solar system

Figure 1: Overview of the morphology-aware translation system: prediction of prepositions, morphological features and generation of inflected forms. German cases: Acc-Accusative, Nom-Nominative, Dat-Dative, Gen-Genitive.

the syntactic function. Other variants contain information of the governing verb/noun, and whether the represented preposition is functional.

For the prediction of prepositions, we combine the following feature types into a linear-chain CRF: *target-side context* (lemmas, POS-tags), *source-side context* (the aligned phrase), *projected source-side information* (relevant target-side words obtained based on source-side parses) and *target-side subcategorizational preferences* (distributional subcategorization information). These features address both functional and content-bearing prepositions, but do not require an explicit distinction between the two categories.

Experiments and Discussion We compare the approach of generating prepositions on the target-side with a morphology-aware SMT system with no special treatment for prepositions. When using "plain" place-holders, there is a considerable drop in BLEU (16.81) in comparison to the baseline (17.38). The annotation of *case* on the place-holders, the best of the abstract representation variants, leads to an improvement (17.23), but still does not surpass the baseline. Additionally, we assess the translation accuracy of prepositions. To allow for an automatic evaluation, we restrict the evaluation to cases where the relevant parts, namely the governing verb and the noun governed by the preposition, are the same in reference and MT output. While there is a minor improvement over the baseline, the difference is very small.

Our approach aims at assigning subcategorized elements to their respective functions and to inflect them accordingly which allows to handle structural differences in source and target language. While the systems fail to improve over the baseline, our experiments show that a meaningful representation of place-holders during translation is a key factor. In particular, the annotation of *case* helps, which can be considered as a "light" semantic annotation. Thus, the addition of more semantically motivated information might lead to a more meaningful representation and remains an interesting idea for future work.

Acknowledgements This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402, the DFG grants *Distributional Approaches to Semantic Relatedness* and *Models of Morphosyntax for Statistical Machine Translation* and a DFG Heisenberg Fellowship.

References

Eneko Agirre, Aitziber Atutxa, Gorka Labaka, Mikel Lersundi, Aingeru Mayor, and Kepa Sarasola. 2009. Use of Rich Linguistic Information to Translate Prepositions and Grammatical Cases to Basque. In *EAMT*.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling Inflection and Word-Formation in SMT. In *EACL*.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying Morphology Generation Models to Machine Translation. In *ACL*.

Marion Weller, Alexander Fraser, and Sabine Schulte im Walde. 2013. Using Subcategorization Knowledge to Improve Case Prediction for Translation to German. In *ACL*.

Marion Weller, Alexander Fraser, and Sabine Schulte im Walde. 2015. Target-Side Generation of Prepositions for SMT. In *EAMT*.