Predicting Prepositions for SMT

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

[1]University of Stuttgart, [2]Ludwig Maximilian University of Munich

1. Introduction & Motivation

Translating prepositions is difficult in SMT

- Convey the source-side meaning
- Meet target-side requirements

How are prepositions generated in translation?

- Functional prepositions: determined by target-side requirements

to believe in sth. → *an etw. glauben*

to learn from sth. → *von jdm. lernen [person]*
aus etw. lernen [abstract]

- Content-bearing prepositions: largely determined by source-side preposition

to sit under/on the table → unter/auf dem Tisch sitzen

- "In-between": source- and target-side play a role

go to the cinema/to the beach → ins Kino/an den Strand gehen

Modeling prepositions on the target-side

- Abstract representation during translation in a morphology-aware EN-DE SMT system
- Generation of prepositions as post-processing step

2. Modeling Prepositions

Subcategorization: difficult to capture in SMT

- Grammatical case of noun phrases corresponds to the syntactic function (subject, direct/indirect object)

Objective: model all subcategorized elements (PP/subject/object) of a verb

- All arguments available in an abstract form
- Are then assigned their respective function
  - overt preposition → PP
  - empty preposition → NP
- Arguments are then inflected accordingly
- Realization of prepositions is independent of structural mismatches of source/target side

to pay attention to sth. → *auf etw. achten*
∅ *etw. beachten*

⇒ both variants are possible, but require a different realization of the preposition depending on the verb

Overview of the translation process

(1) Building the morphology-aware SMT-system
- lemmatized representation for translation
- target-side prepositions are replaced with place-holders; "empty" place-holders are inserted at the beginning of NPs
- "empty" prepositions added to source-side NPs

(2) Generating surface forms
- prediction and realization of place-holder prepositions as overt preposition (PP) or "empty" preposition (NP)
- prediction of inflection-relevant morphological features
- generation of inflected forms

3. Overview: Translation and Prediction Steps

input	lemmatized SMT output	prep.	morph. features	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

German Cases: *Nominative* – subject; *Accusative* – direct object; *Dative* – indirect object; *Genitive* – nominal modifier

4. Features for Predicting Prepositions

- **Target-side context**: adjacent lemmas+POS tags
- **Source-side features**
  - aligned word on source-side: overt or empty preposition
  - governed noun and its syntactic function to its governor
  - governing verb or noun of source-side preposition
- **Projected source-side features**
  - governing target verb, governed target noun
- **Distributional subcategorization preferences**
  - information in form of e.g. *verb-preposition-case* tuples
  - learn, whether a given combination predominantly occurs as subject, direct/indirect object, PP or noun-noun modification
- Prediction models: CRFs trained with *Wapiti*
  Prediction accuracy: 73.5% (prep+case); 85.7% (prep)

6. Abstract Representation of Prepositions

"Basic" place-holder → decreased translation quality

S1 Plain place-holders

Enriched abstract representation

S2 Grammatical case
- overt preposition: case often indicator of content (direction,location)
- empty preposition: case indicates the syntactic function

S3 Governor of the preposition (verb or noun)

S4 Functional vs. content-conveying
- subcategorization lexicon: is a preposition in a given context functional?

S5 Assuming that functional prepositions convey less in terms of meaning
- replace functional prepositions with place-holders
- keep "regular" prepositions for content-conveying prepositions

5. Prediction Features in the Training Data

lemma	gloss	source-side			projected source-side		target-side			label
		prp	func,noun	g.verb	noun	g.verb	subcat			
aber	*but*	–	–	–	–	–		–		-
PREP	*PRP*	∅	subj, we	endure	wir	leiden	∅-Nom:5	∅-Acc:0	*unter-Dat*:4	∅-Nom
wir	*we*	–	–	–	–	–		–		Nom
leiden	*suffer*	–	–	–	–	–		–		-
…	…	…	…	…	…	…		…		…
auch	*too*	–	–	–	–	–		–		-
PREP	*PRP*	∅	obj, effect	endure	Treibhauseffekt	leiden	∅-Nom:5	∅-Acc:0	*unter-Dat*:4	**unter-Dat**
die	*the*	–	–	–	–	–		–		Dat
Treibhauseffekt	*greenhouse effect*	–	–	–	–	–		–		Dat

Source sentence with inserted empty prepositions: … , ∅ *we too are having to endure ∅ the greenhous effects*

7. Experiments

- Standard phrase-based Moses system
- 4.3M parallel EN–DE sentences, 10.3M lines LM-data
- Test/tuning sets: 3000 sentences news data

System	Prepositions	BLEU	CRF
Baseline$_{surface}$	–	16.84	–
Baseline$_{morphology}$	–	17.38	–

	Representation of place-holders	BLEU source	BLEU src+sub
S1	□	16.81	16.77
S2	□+Case	17.23	17.23
S3	□+Case+(V|N)	16.91	16.89
S4	□+Case+(V|N)+subcat	17.09	17.08
S5a	□+Case+(V|N): functional prp+Case+(V|N): non-func.	17.12	17.06
S5b	□+Case+(V|N): functional prp+Case+(V|N): non-func.	17.29	17.29

- No improvement over baseline; best result obtained with annotation of case (S2)

Automatic evaluation of generated prepositions

- Subset where relevant parts (governed noun, governing verb) match with the reference
- No real improvement over baseline

Example

SRC	malmon 's team will have to **improve** on recent performances .
BL	malmon das Team wird über die jüngsten Leistungen zu **verbessern**. *malmon the team will over the recent performances improve.*
NEW	malmon das Team hat ∅ die jüngsten Leistungen zu **verbessern** . *malmon the team has-to ∅ the recent performances improve*
REF	muss sich das Malmon-Team im Vergleich zu den vergangenen Auftritten … steigern *must -refl- the malmon-team in comparison to the past performances … improve*

8. Conclusion & Future Work

- Generation of prepositions based on an abstract representation using source and target features

⇒ handle structural differences between source/target-side

- No improvement over morphology-aware baseline
  - annotation of grammatical case → best system

How to improve the current method?

- Abstract representation
  - grammatical case: light semantic annotation
  - obtain a more meaningful representation by more semantically motivated annotation to represent the class of a preposition (*temporal, local, directional, …*)
- Integration of the generation step
  - integrate into decoding process ("synthetic phrases", Chahuneau et al. 2013)

⇒ generation of prepositions appropriate for respective context, but translation without place-holder representation

9. Selected Related Work

Agirre et al. (2009): *Use of Rich Linguistic Information to Translate Prepositions and Grammatical Cases to Basque*. In Proceedings of EAMT.

Weller et al. (2015): *Target-side Generation of Prepositions for SMT*. In Proceedings of EAMT.

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