Using Subcategorization Information to Improve Case Prediction for Translation to German

Marion Weller[1], Alexander Fraser[2], Sabine Schulte im Walde[1]
[1]University of Stuttgart, [2]Ludwig Maximilian University of Munich

1. Introduction

We present an English-German SMT system that deals with complex target-side morphology by applying a **two-step translation process**: (cf. [2],[3])

- translation model built on stems;
- prediction of morphological features, generation of inflected forms.

Improving *case* prediction

- Due to the flexible German clause ordering, case is difficult to predict.
- Case is an important indicator of the role of an NP in the sentence; the most difficult is to distinguish
  - syntactic functions (subject, direct/indirect object)
  - modifying NPs (genitive modification).
- **New features for case prediction**
  - projection of source-side syntactic information;
  - information about target-side syntactic frames obtained from dependency-parsed corpora.

2. Overview of the inflection process

Morphological Features

- The gender of an NP is part of the stem.
- English input determines the number of an NP.
- Strong/weak inflection depends on the choice of determiner and the setting of the other features.
- There are 4 values for case: *nominative* (Subject), *accusative* (direct object), *dative* (indirect object) and *genitive* (modification, object in rare cases).

Feature prediction and inflection

- Individual sequence models for each morph. feature
- The models have access to stems, POS-tags within a window of four positions
- Generate inflected forms using features and stems:
  `blau<ADJ><nom><fem><sg><weak>` → `blaue` (cf. [1])

SMT output	predicted features	inflected forms	gloss
solche<D>	Masc.Nom.Pl.St	**solche**	such
Bus<N><M><Pl>	Masc.Nom.Pl.Wk	**Busse**	buses
haben<VAFIN>	–	haben	have
Zugang<N><M><Sg>	Masc.Acc.Sg.St	**Zugang**	access
zu<APPR><Dat>	–	zu	to
die<D><Def>	Neut.Dat.Sg.St	**dem**	the
Land<N><M><Sg>	Neut.Dat.Sg.Wk	**Land**	country

Table: Processing steps for the input sentence *these buses may have access to that country.* (simple case prediction, cf. [3])

3. Motivation for case modeling

Source-side features

(1) dass er$_{NOM}$ den Minister$_{ACC}$ unterstützt
that he$_{NOM}$ supports the minister$_{ACC}$

(2) dass ihn$_{ACC}$ der Minister$_{NOM}$ unterstützt
that the minister$_{NOM}$ supports him$_{ACC}$

- *Minister* (*minister*) is a plausible subject and direct object for the verb *unterstützen* (*support*).
- Projecting the NP's roles from the input sentence helps to disambiguate the syntactic function.

Subcategorization information

(1) Der Chef$_{NOM}$ gab den Bericht$_{ACC}$ dem Mitarbeiter$_{DAT}$
The boss$_{NOM}$ gave his colleague$_{DAT}$ the report$_{ACC}$

(2) Der Chef$_{NOM}$ stimmte dem Bericht$_{DAT}$ des Kollegen$_{GEN}$ zu
The boss$_{NOM}$ agreed on the report$_{PP}$ of his colleague$_{PP}$

- *geben* (*give*) has a bias for a ditransitive subcategorization frame: subject, benefactive, patient.
- Bericht (*report*) is more likely to be patient (direct object) than Mitarbeiter (*employee*).
- *zustimmen* (*agree*) has a preference for only selecting subject and indirect object theme.
- The remaining NP cannot receive case from the verb and is thus a genitive modifier of the NP$_{DAT}$.

4. Subcategorization features

input	stemmed output	inflected	gloss
the	d	**die**	*the*
government	Regierung	Regierung	*government*
threatens			
the	d	**der**	*of the*
united	vereinigt	vereinigt**en**	*united*
states	Staat	Staat**en**$_{GEN}$	*states*
	droht	droht	*threatens*

Table: Example for case confusion in SMT-output.

Subcategorization information

- External knowledge base comprises dependency information on verbal subcategorization frames.
- We model the association strength for verb-noun pairs and for N-N$_{gen}$ pairs.
- Subcategorization information is obtained from news data (200M words) and Europarl.

Verb-noun tuples with case information

tuple	gloss	case values nom	acc	dat
Erfahrung gewinnen	*gain experience*	38	242	2
Erfahrung zeigen	*experience show(s)*	4708	412	6
Erfahrung entsprechen	*correspond to experience*	45	9	201

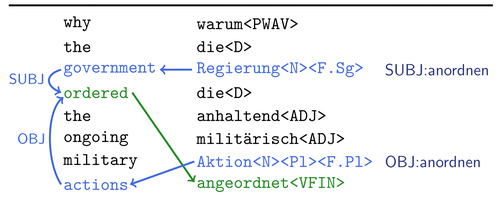
N-N$_{gen}$ tuples with frequency information

tuple	gloss	frequency
Ergebnis Wahl	*result of vote*	449
Entwicklung Produkt	*development of product*	814

5. Integration of source-side and subcategorization features

Integration of source-side features

- English dependency relations are transferred to the SMT output based on the word alignment.
- Information about the complete tuple (verb+noun and N-N$_{gen}$) is annotated as bigram, e.g. `Regierung+anordnen`.



Integration of subcategorization information

- Extraction of verb-noun tuples and candidates for N-N$_{gen}$ constructions:
  - based on syntactic trees produced by a hierarchical SMT-system
  - derived from source-side dependencies via word alignment
- Look up co-occurrence probabilities/frequencies.

stems	gloss	Acc	Dat	Nom	verb	Gen	N1	gold
Unternehmen<N>	*companies*	0.06	0.00	0.94	erhalten	–	–	*Nom*
sollten<VFIN>	*should*	–	–	–	–	–	–	–
finanziell<A>	*financial*	–	–	–	–	–	–	*Acc*
Mittel<N>	*funding*	1.00	0.00	0.00	erhalten	–	–	*Acc*
für<APPR><Acc>	*for*	–	–	–	–	–	–	–
d<ART>	*the*	–	–	–	–	–	–	*Acc*
Einführung<N>	*introduction*	–	–	–	–	–	–	*Acc*
neu<ADJ>	*new*	–	–	–	–	–	–	*Gen*
Technologie<N>	*technologies*	0.00	0.00	0.00	–	100	Einführung<N>	*Gen*
erhalten<VINF>	*obtain*	–	–	–	–	–	–	–

6. Experiments and evaluation

0	1	2	3	4
surface system	simple prediction	subcat. features	source-side features	source-side + subcat. features
13.43	14.02	14.05	14.10	14.17

Table: BLEU scores for different inflections (1-4).

- Hierarchical SMT system using GHKM target-side syntax trained on WMT-2009 data (Europarl)
- Inflection prediction better than surface system.
- Systems 1-4: different inflections of the same SMT output; system 1 does not use new features.
- No significant difference between the enriched systems and the simple prediction system.
- No changes in stem sequence, but different inflection; BLEU can hardly capture the difference:
  [*den vereinigten Staaten*]$_{ACC}$ (*the United States*)
  [*der vereinigten Staaten*]$_{GEN}$ (*of the United States*)

Manual evaluation

- Human annotators prefer the enriched system.

	system 4 preferred	system 1 preferred	equal
Person 1	23	5	18
Person 2	21	11	14
Person 3	29	8	9

Table: Manual evaluation: simple (1) vs. both features (4).

7. Conclusion

- We presented a two-step SMT system that translates into stems and generates inflected forms.
- We illustrated the need for external knowledge sources to model case and presented a translation system using source-side syntactic features and a subcategorization database.
- A manual evaluation showed that the proposed features have a positive impact.
- First integration of explicit subcat-information from large monolingual corpora into SMT.

8. References

[1] H. Schmid, A. Fitschen, U. Heid. SMOR: a German Computational Morphology covering Derivation, Composition and Inflection. *LREC 2004.*
[2] K. Toutanova, H. Suzuki, A. Ruopp. Applying Morphology Generation Models to Machine Translation. *ACL-HLT 2008.*
[3] A. Fraser, M. Weller, A.Cahill, F. Cap. Modeling inflection and Word Formation in SMT. *EACL 2012.*

This work was funded by the DFG Research Grants *Distributional Approaches to Semantic Relatedness* and *Models of Morphosyntax for Statistical Machine Translation* and the DFG Heisenberg Fellowship SCHU-2580/1-1.