Using Noun Class Information to Model Selectional Preferences for Translating Prepositions in SMT

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

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

1. Motivation & Introduction

- Translating prepositions is difficult in SMT

⇒ Reproduce the preposition's meaning in the input sentence

⇒ Take into account target-side context

- Some prepositions convey a meaning → straightforward translation
  - *to sit* UNDER/ON *the table*
- Some prepositions are functional → largely depend on target-language restrictions
  - *to believe* IN *something*
- Prepositions are typically determined by a governor
  - Verbs: *to believe* IN *sth.*
  - Nouns: *an interest* IN *sth.*
- Additionally, prepositions can depend on the class of nouns that are governed
  - *to learn* FROM *[a person]* → *lernen* VON *[einer Person]*
  - *to learn* FROM *[the past]* → *lernen* AUS *[der Vergangenheit]*
- We enrich an EN-DE string-to-tree SMT system with noun class information to model selectional preferences

2. Modeling Selectional Preferences

Methodology

- Annotating noun class information into the parse trees used to train a syntax-based SMT system

`to learn from [NP]` → `lernen von [NP`$_{Person}$`]`
→ `lernen aus [NP`$_{Abstract}$`]`

- The enriched translation rules are restricted to those of a specific semantic class appropriate for a given context
- Using noun class information to obtain more precise translation rules that incorporate selectional preferences
- Aims at introducing a semantic level into SMT

Noun class information

- Three variants of noun class information
  1. Classes induced from the lexical resource GermaNet
     ⇒ Conceptually refined target-language information
  2. Cluster analyses based on window information
  3. Cluster analyses based on syntactic features
     ⇒ Generalization over contexts (in "raw" form or based on syntactic structures): take into account more target-language information
- Comparing resource-based and distributional information

PP rule generation

- Semantically fine-grained information might lead to a loss of generalization
- Make generic rules accessible to the system
- Generate new PP rules that are not accessible to the baseline in order to cover all functional prepositions

3. Obtaining Noun Class Information

- Pre-processing: identification of named entities
  - Consistent distinction of named entities and common nouns
  - Named entities classified into *organization, location, person, rest*
- Pre-processing: compound handling
  - Noun compounding is very productive: compound-splitting
  - Compounds are added into classes based on their head nouns

GermaNet

- Lexical-semantic taxonomy that groups words of the same concept into synsets (→ WordNet)
- Look up the GermaNet class for a given hierarchical level

Clustering

- Standard k-Means implementation in R
- What number of clusters provides
  - A good representation of the nouns
  - An optimal level of abstraction for the SMT system?
- Varying cluster sizes: 10 - 300 clusters
- Clustering based on window information
  - Content words from a window of 10 words to each side of the noun

⇒ Often results in "topic-like" clusters

- Clustering based on syntactically-motivated features
  - Prepositions governing the target nouns **P**
  - Verbs subcategorizing the target nouns **VO**
  - Verbs governing the target nouns in a prepositional phrase **VPN**
  - Nouns governing the target nouns in a prepositional phrase **NPN**

⇒ Using subcategorization criteria aims at obtaining classes that provide salient information for modeling the choice of prepositions

4. Annotation of PP-nodes and NP-nodes of the Target-Side Parse Trees

```
<tree=s>
  <tree=adjd> wirtschaftlich </tree>
  <tree=vafin-haben> hat </tree>
  <tree=np-LOC>
    <tree=ne-LOC> malaysia </tree>
  </tree>
  <tree=vp>
    <tree=pp-von-167>
      <tree=prp-von-167> von </tree>
      <tree=pposat> seinen </tree>
      <tree=nn-167> nachbarn </tree>
    </tree>
    <tree =vvpp> gelernt </tree>
  </tree>
</tree>
```

*economically, Malaysia has **learned from its neighbors**.*

```
<tree=s>
  <tree=kous> dass </tree>
  <tree=np-180>
    <tree=art> die </tree>
    <tree=nn-180> amerikaner </tree>
  </tree>
  <tree=vp>
    <tree=pp-aus-291>
      <tree=prp-aus-291> aus </tree>
      <tree =art> der </tree>
      <tree=nn-291> vergangenheit </tree>
    </tree>
    <tree=vvpp> gelernt </tree>
  </tree>
  <tree=vafin-haben> hätten </tree>
</tree>
```

*that the Americans had **learned from the past**.*

5. Using Noun Class Information in SMT

- Create two variants for the translation of learned from NN

```
VP → PP-von-167 gelernt
VP → PP-aus-291 gelernt
```

- Nouns of the classes 167 (*person*) and 291 (*abstract concept*) are appropriate fillers for the PPs

Back-off strategies

- Add baseline rules (rules without annotation) **BL**
- No annotation for rules based on low-frequency source-target pairs ($f \leq 5$) **BL+cutoff**

Generating new PP rules

- Not all potentially necessary rules might be available
- Provide the full possible set of rules containing functional prepositions (i.e. prepositions with little or no meaning)
- Create new rules for a set of 17 functional prepositions
- Translation probabilities for new rules: based on co-occurrence frequencies extracted from large corpora

original rule (target-side)	prob.
VP → [pp-von-166] lernen , [s]	1

new PP rules (target-side)	prob.
VP → [pp-aus-166] lernen , [s]	0.159
VP → [pp-für-166] lernen , [s]	0.021
VP → [pp-in-166] lernen , [s]	0.126
VP → [pp-mit-166] lernen , [s]	0.021
VP → [pp-von-166] lernen , [s]	0.336
VP → [pp-über-166] lernen , [s]	0.336

pnv-tuple	freq
aus nn-166 lernen	38
für nn-166 lernen	5
in nn-166 lernen	30
mit nn-166 lernen	5
von nn-166 lernen	80
über nn-166 lernen	80

- Adding generated rules **new rules**
- Adding both back-off and new rules **BO+new**

6. Experiments

- String-to-tree Moses system with GHKM extraction
- Morphology-aware translation system allows to explicitly model portmanteau prepositions
- SMT system: 1.5M parallel sentences (Europarl + news)
- Feature extraction: additional 44M sentences (web data)

System	BLEU
Baseline	13.95
GermaNet-2 (25)	13.93
GermaNet-3 (79)	13.77
GermaNet-4 (175)	13.67
GermaNet-5 (392)	13.67

System	BLEU
Window10	14.01
Window50	14.18
Window75	13.69
Window100	14.13
Window300	13.71

Syntactic features	P	VO	VPN	NPN
100 classes	13.85	13.85	13.79	13.71
50 classes	13.84	14.06	14.06	13.91

System	BL+cutoff	BL	new rules	BL+new rules
Window50	13.95	13.99	14.11	13.98
Window75	14.16	13.96	13.66	14.01
Window100	14.01	13.94	14.14	14.02

7. Discussion & Conclusion

- None of the systems is better than the baseline
- Manual evaluation of correctly translated prepositions: little difference between the systems
- No systematic behaviour or types of prepositions that are translated better or worse across the systems
- Particularly difficult: prepositions with a predominant literal meaning in an infrequent subcategorized context

EN *for example, Germany has been criticized **for** passivity*
DE *beispielsweise, Deutschland *für Passivität kritisiert* worden
REF *wegen Passivität wurde zum Beispiel Deutschland kritisiert*

- Context-dependent interaction of being a functional or content-bearing preposition, importance of involved noun classes: not well-captured by inflexible annotation method
- Noun class annotation into parse trees → hard constraint
- Compensate for overly specific rules with non-annotated rules and rules synthesized from monolingual data
- No generally applicable level of semantic information: rigid annotation → rules of the same degree of specificity
- Results demonstrate that
  - Clustering based on window co-occurrence seems to be more robust than syntax-based clusters or GermaNet ⇒ **Resources**
  - Parse tree annotation is not flexible enough to take into account varying needs of different contexts ⇒ **Integration method**
- Idea for future work: combine distributional (→ *robust, coverage*) and resource-based information (→ *high quality*) to obtain salient information on selectional preferences

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.