Modelling inflection for SMT into Finnish

Fabienne Cap, Sara Stymne, Jenna Kanerva, Marion Di Marco, Filip Ginter, Jörg Tiedemann

Uppsala, Turku, Stuttgart and Helsinki

Motivation: SMT into Finnish

English	Finnish	Case
the house	talo	nominative
of the house	talon	genitive
the house	talo	accusative
part of the house	taloa	partitive
in the house	talossa	inessive
out of the house	talosta	elative
into the house	taloon	illative
at the house	talolla	adessive
away from the house	talolta	ablative
to the house	talolle	allative
as the house	talona	essive
to become the house	taloksi	translative
without the house	talotta	abessive
with houses	talоin	instructive
with (all his) houses	taloineen	komitative

Inspiration

- System based on Fraser, Weller-Di Marco, Cap, Cahill Morphological inflection and word formation for **German**
  - Standard phrase-based SMT on modified data
  - Preprocessing: "lemmatize" the input and split compounds

    die<+ART><Def> Tür<+NN><Fem><Sg>
    ist<+V> blau<+ADJ><Pos>
  - Post-processing:

    Predict inflection features using CRF
    Generate inflected forms using finite-state technology
    Merge compounds
- Adapted to **Finnish** in this work

Underspecified Representation in German

English	German	Analysed	Reduced
The	Die	die<+ART><Def><**Fem**><**Nom**><**Sg**><**St**>	die<+ART><Def>
door	Tür	Tür<+NN><Fem><**Nom**><Sg>	Tür<+NN><Fem><Sg>
is	ist	sein<+V><**3**><**Sg**><**Pres**><**Ind**>	ist<+V>
blue	blau	blau<+ADJ><Pos><**Pred**>	blau<+ADJ><Pos>
.	.	.<+PUNCT><Norm>	.<+PUNCT><Norm>

- Use a rule based morphological analyser
- Remove **case** from nouns
- Remove **gender**, **gender**, **number** and **strength** from adjectives and determiners
- Leave morphological features that are inherent to the word

Translation Step, German

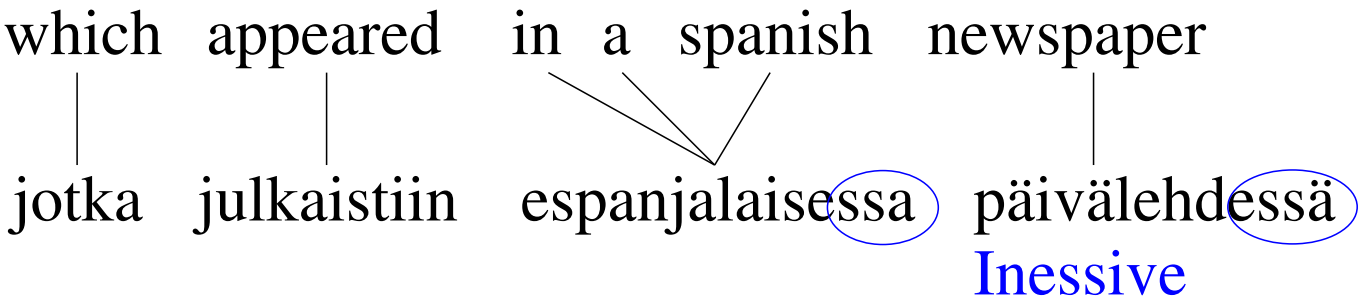
English	→	**Underspecified**	Generated
the	→	die<+ART><**Def**><Neut><Sg><Nom>	d**as**
blue	→	blau<+ADJ><Pos><Neut><Sg><Wk><Nom>	blau**e**
house	→	Haus<+NN><**Neut**><**Sg**><Nom>	Haus
with	→	mit<+PREP><**Dat**>	mit
the	→	die<+ART><**Def**><Fem><Sg><Dat>	d**er**
red	→	rot<+ADJ><Pos><Fem><Sg><Wk><Dat>	blau**en**
door	→	Tür<+NN><**Fem**><**Sg**><Dat>	Tür

German vs. Finnish

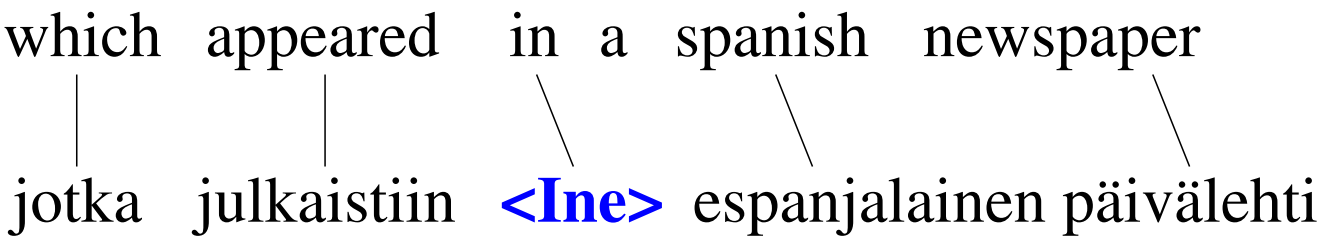
Feature	German	Finnish
adjective noun agreement	✓	✓
inflects for gender	✓	✗
strong/weak inflection	✓	✗
use of articles	✓	✗
standalone prepositions	✓	✗
#cases	4	15 (!)

In the present work,
we predict **cases** on Finnish **adjectives** and **nouns**
Use placeholder prepositions
Split and merge compounds

Placeholder Prepositions



Placeholder Prepositions



Advantages of placeholder prepositions:

- better alignment of prepositions
- help CRF case prediction

Feature Prediction with CRFs

Finnish is challenging:

- 12-16 labels to assign
- not tractable on Europarl corpus
- instead: 1/4 of Europarl

Clean data re-prediction results (on devset):

placeholders	**precision**
no	69.17%
all	86.51%
some	83.59%

(For German: 94.29% accurracy)

Goals

- improve SMT quality
- reduce data sparsity
- better word alignment
- more coherent and fluent sequences
- produce unseen inflectional variants
- create standalone prepositions in order to improve word alignment to English, and thus improve SMT quality

SMT Experiments

- Data taken from WMT 2015
  - Training: Europarl
  - LM: Europarl
  - Dev/Test: News
- Preprocessing:
  - OmorFi rule-based morphological analyser
    - Data-driven method as back-off
  - Turku Finnish dependency parser
- Moses phrase-based SMT system
- Tuning of feature weights against lemmatised reference
- No compound processing in this variant

SMT Results

system	BLEU
baseline	9.60
no prepositions	9.39
all prepositions	9.74
some prepositions	9.89

SMT results

- Modest improvements on Bleu
- Manual evaluation of 100 sentences:
  - Baseline preferred: 24
  - Inflection system preferred: 41
- Around 300 novel word forms

Sample of merged compounds

From system with added compound processing:

kokonaisnousu total increase	kokonainen nousu total increase
ulkomaalaisia foreigner	ulko maalainen outside country-from
verensokeritasoa blood sugar level	veren sokeri taso blood (Gen) sugar level
suomalaiselokuva Finnish movie	suomalainen elo kuva Finnish living picture
ihmissalakuljetusbisnekseen people smuggling business	ihminen sala kuljetus bisnes human secret transportation business
maailmanmestaruuskilpailuissa world championship competition	maa ilman mestaruus kilpailu ground air championship competition (maailma / world)

Questions for the future

- Can these challenges be better solved by NMT than SMT?
- Can any of these techniques be useful for NMT?
  - Placeholder prepositions
  - Lemma representations
  - Compound processing
  - Case prediction
- Is there still a place for SMT?
  - Are these techniques worth persuing?