

# Addressing Problems across Linguistic Levels in SMT: Combining Approaches to Model Morphology, Syntax and Lexical Choice

Marion Weller-Di Marco<sup>1,2</sup> Alexander Fraser<sup>2</sup> Sabine Schulte im Walde<sup>1</sup>  
<sup>1</sup>IMS – University of Stuttgart    <sup>2</sup>CIS – Ludwig-Maximilians-University of Munich  
{dimarco|schulte}@ims.uni-stuttgart.de    fraser@cis.lmu.de

## PROBLEMS ACROSS LINGUISTIC LEVELS

### Structural differences in source/target language

- Often difficult to capture with **word alignment**
- **Long-distance reordering** is typically **costly** during translation

### Lexical choice

- **Word sense disambiguation**, **selectional preferences**, ...
- Translation of **multi-word expressions**

### Morphological complexity

- **Data sparsity** due to uncovered inflected forms
- Difficulty to produce the **correct target-side inflection** based on available information

## COMBINING APPROACHES

### • Pre-processing – syntactic level

Source-side reordering (Gojun and Fraser, 2012)

### • At decoding time – lexical level

Discriminative classifier to score translation rules using source-side context (Tamchyna et al., 2014)

### • Post-processing – morphological level

Target-side inflection prediction (Fraser et al. 2012)

→ EN-DE phrase-based SMT system

→ Do individual gains add up when combining approaches?

→ Are there side-effects between the linguistic levels?

## SYNTACTIC LEVEL

- English **verbs** are moved to the **expected German position**
  - moving the verb to *verb-final position*
  - moving the verb to *verb-second position*
- Resulting structure considerably different from “regular” English

## LEXICAL LEVEL

- **Sentence-level source-side features**, target features restricted to phrase
- Features for **discriminative model**
  - pos/word/lemma (source-side window, target-side phrase)
  - dependency information (source-side)

## MORPHOLOGICAL LEVEL

- **Inflection prediction** process to handle **nominal morphology**:
  - translation in stemmed representation
  - generation of inflected forms
- Markup with translation relevant inflectional features

## EFFECTS OF SOURCE-SIDE REORDERING

### Reordering separates the verb and its arguments

that the **ground** **was** permanently frozen    that the **ground** permanently frozen **was**  
|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |  
dass der **boden** **ständig** gefroren **war**    dass der **boden** ständig gefroren **war**

→ **Negative effect on verb-subject agreement** Ramm et al. (2016)

in the current crisis , the us federal reserve and the european central bank **cut** interest rates  
|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |  
in der aktuellen krise **senken** die us-notenbank und die europäische zentralbank die **zinssätze**  
in the current crisis , **cut** the us federal reserve and the european central bank **interest rates**  
|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |  
in der aktuellen krise **senken** die us-notenbank und die europäische zentralbank die **zinssätze**

→ **Assumption: loss of context for support-verb constructions**

*cut* → *senken*    ‘decrease’    **non-literal meaning**  
*cut* → *schneiden*    ‘cut (with knife)’    **literal meaning**

## EXTENDED SOURCE CONTEXT

### Number and tense features

- **Number** to model **subject-verb agreement**
  - number of verbs is not always obvious: *went, could, will*
- **Auxiliaries** in compound tenses
  - annotate *past/non-past* for compound tenses
  - annotate main verb for auxiliary choice in past tense

### Support verb constructions

- SVC: **verb + predicative noun** (e.g. *make a contribution*)  
the verb does not contribute its full meaning
- Annotation of support verb status can improve SMT  
→ **literal/non-literal verb translation** Cap et al. 2015
- Add explicit SVC information
  - annotate the **degree of association** between noun and verb
  - **mark as SVC** if association strength above a threshold

## RESULTS

- EN-DE phrase-based Moses system
- 4,5 M parallel sentences
- 5-gram language model (45 M sentences)

### Morpho-syntactic + lexical strategies

system	basic	VW-1 pos/lem	VW-2 pos/lem/dep
Surface	19.45	19.81	19.90
Surface V-Reordered	19.71	20.24	20.27
MorphSys	19.81	19.80	19.93
MorphSys V-Reordered	20.08	20.51	20.50

### Annotating number + tense information

system	VW-2	VW-2 +num	VW-2 +num+tense
MorphSys	19.93	20.00	20.02
MorphSys V-Reordered	20.50	20.57	20.62*

### Annotating support verb information

- SVC information already in dependencies
- Explicit annotation: no improvement

## EXAMPLES FOR NUMBER AND TENSE FEATURES

<b>SRC</b>	i really feel that <b>he should</b> follow in the footsteps of the other guys .
<b>reordered</b>	i really feel that <b>he</b> in the footsteps of the other guys follow <b>should</b> .
<b>VW2</b>	ich bin wirklich der Meinung , dass <b>er</b> in die Fußstapfen der anderen Jungs folgen <b>sollten<sub>PL</sub></b> . <i>i am really of-the opinion , that <b>he</b> in the footsteps of the other guys follow <b>should</b> .</i>
<b>+NumTense</b>	ich bin wirklich der Meinung , dass <b>er</b> in die Fußstapfen der anderen Jungs folgen <b>sollte<sub>SG</sub></b> . <i>i am really of-the opinion , that <b>he</b> in the footsteps of the other guys follow <b>should</b> .</i>

<b>SRC</b>	it <b>would</b> thus <b>be</b> suitable to assist illegal immigration into the usa .
<b>reordered</b>	it <b>would</b> thus suitable <b>be</b> illegal immigration into the usa to assist .
<b>VW2</b>	es <b>wäre</b> daher geeignet <b>sein</b> , die illegale Einwanderung in die USA zu unterstützen . <i>it <b>would-be</b> thus suitable <b>be</b> , the illegal immigration into the usa to assist .</i>
<b>+NumTense</b>	es <b>wäre</b> daher ideal , illegale Einwanderung in die USA zu unterstützen . <i>it <b>would-be</b> thus ideal , illegal immigration into the usa to assist .</i>

## CONCLUSION

- **Combination of established approaches** to address the three linguistic levels *Morphology, Syntax and Lexical Choice*
- The strategies are **complementary**
- **Reordered systems** benefit more from **discriminative model**
- Additional features aiming at **verbal inflection** lead to **further improvement**