Deepfix:

Statistical Post-editing of Statistical Machine Translation Using Deep Syntactic Analysis

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

ACL SRW, Sofia, 6th August 2013
Source text in English:

EU criticizes not only the Greek government.
Src: EU criticizes not only the Greek government

Translation:
EU kritizuje nejen řecká vláda
Motivation

- Source text in English:

EU criticizes not only the Greek government

- Google Translate to Czech (6\textsuperscript{th} Aug 2013):

EU kritizuje nejen \textit{řecká vláda} nominative (subject)

- Not only the Greek government criticizes EU
Motivation

- Source text in English:
  
  EU criticizes not only the Greek government

- Google Translate to Czech (6\textsuperscript{th} Aug 2013):
  
  EU kritizuje nejen řecká vláda

  - Not only the Greek government criticizes EU

- Post-editation by Deepfix:
  
  EU kritizuje nejen řecko\textsubscript{u} vládu

  - EU criticizes not only the Greek government
Outline

1. Problem definition
2. Sentence analysis
3. Sentence post-editing
4. Results
Outline

1. Problem definition
   - Errors in valency in SMT outputs
2. Sentence analysis
3. Sentence post-editing
4. Results
Outline

1. Problem definition
   - Errors in valency in SMT outputs
2. Sentence analysis (DEEP)
   - Deep dependency parsing
3. Sentence post-editing
4. Results
Outline

1. Problem definition
   - Errors in valency in SMT outputs

2. Sentence analysis (DEEP)
   - Deep dependency parsing

3. Sentence post-editing (FIX)
   - Statistical model of valency

4. Results
1. Problem definition
   - Errors in valency in SMT outputs

2. Sentence analysis (DEEP)
   - Deep dependency parsing

3. Sentence post-editing (FIX)
   - Statistical model of valency

4. Results
   - Automatic & manual evaluation of Deepfix
Subject – object dichotomy

- English: **position** (left/right constituent)
  - Subject *criticize* Object

- Czech: **morphological case** (nominative/other); word order relatively free
  - Subject _nominative_ *kritizovat* Object _accusative_
  - Object _accusative_ *kritizovat* Subject _nominative_
  - Subject _nominative_ Object _accusative_ *kritizovat*
  - Object _accusative_ Subject _nominative_ *kritizovat*
Valency of criticize (kritizovat)

- example sentence
  - EU \text{subject} criticizes not only the Greek government \text{object}
  - EU \text{nominative} kritizuje nejen řeckou vládu \text{accusative}
Valency of *criticize* (*kritizovat*)

- **example sentence**
  - *EU*_{subject} **criticizes** not only the *Greek government*_{object}
  - *EU*_{nominative} **kritizuje** nejen řeckou vládu_{accusative}

- **a valency frame of a verb**
  - subject **criticize** object
  - nominative **kritizovat** accusative
Valency of criticize (kritizovat)

- example sentence
  - $EU_{subject}$ criticize $not$ only the $Greek$ government $object$
  - $EU_{nominative}$ kritizuje $nejen$ řeckou vládu $accusative$

- a valency frame of a verb
  - subject criticize object (position)
  - nominative kritizovat accusative (cases)
Valency of *criticize* (*kritizovat*)

- **example sentence**
  - *EU* criticizes not only the Greek government
  - *EU* kritizuje nejen řeckou vládu

- **a valency frame of a verb**
  - subject *criticize* object (position)
  - nominative *kritizovat* accusative (cases)

- **decomposition into head-argument pairs**
  - *(to criticize, government) ~ (kritizovat, vládu)*
Sentence analysis (DEEP)

- tagging & lemmatization
  - combination of rule-based and statistical approach
- word-alignment
  - unsupervised methods (Giza++)
- dependency parsing
  - statistical, trained on manually created treebanks
  - parser adapted for parsing of SMT outputs
- induction of deep structure (tectogrammar)
  - rule-based
Deep syntactic dependency trees

EU criticizes the Greek government

EU kritizuje řecká vláda
Deep syntactic dependency trees

EU criticizes the Greek government

EU kritizuje řecká vláda

EU
criticize
government
GREK

ŘECKÁ
vláda

EU
kritizovat

Greek
Deep syntactic dependency trees

EU criticizes the Greek government

EU kritizuje řecká vláda
(head, arg) pair identification

EU criticizes
the Greek government

EU kritizuje
řecká vláda

Criticize
V, predicate

EU
N, subject

Government
N, object

Greek
A, attribute

EU
N, nominative

Vláda
N, nominative

Rècká
A, attribute
Valency models (FIX)

- $P(\text{arg}_{\text{case}} | \text{head}_{\text{lemma}}, \text{English}_{\text{arg}_{\text{case}}})$
- $P(\text{arg}_{\text{case}} | \text{head}_{\text{lemma}}, \text{English}_{\text{arg}_{\text{case}}}, \text{arg}_{\text{lemma}})$
- estimated from CzEng 1.0 (15M parallel stcs)
Argument case probabilities

- $P(\text{nominative} \mid kritizovat, \text{object}) = 0.03$
- $P(\text{accusative} \mid kritizovat, \text{object}) = 0.80$
Argument case probabilities

- $P(\text{nominative} \mid \text{kritizovat}, \text{object}) = 0.03$
- $P(\text{accusative} \mid \text{kritizovat}, \text{object}) = 0.80$
- threshold: 0.55
Argument case correction

- $P(\text{nominative} \mid \text{kritizovat}, \text{object}) = 0.03$
- $P(\text{accusative} \mid \text{kritizovat}, \text{object}) = 0.80$
- threshold: 0.55
Sentence correction

- Statistical machine translation output:
  
  EU kritizuje nejen řecká nominative vláda nominative

- Not only the Greek government criticizes EU
Sentence correction

- Statistical machine translation output:
  
  \[ EU \text{ kritizuje nejen } \acute{\text{r}}\text{ecká} \text{nominative } \text{vláda} \text{nominative} \]
  
  - Not only \textit{the Greek government} criticizes \textit{EU}

- Valency model correction:
  
  \[ EU \text{ kritizuje nejen } \acute{\text{r}}\text{ecká} \text{nominative } \text{vládu} \text{accusative} \]
Sentence correction

- Statistical machine translation output:
  
  *EU kritizuje nejen řeckánominative vláda nominative*

  - Not only *the Greek government* criticizes EU

- Valency model correction:
  
  *EU kritizuje nejen řecká nominative vládu accusative*

- Agreement enforcement:
  
  *EU kritizuje nejen řeckou accusative vládu accusative*

  - *EU criticizes not only the Greek government*
Automatic evaluation (BLEU)

| Dataset   | SMT output | after Deepfix |
|-----------|------------|---------------|
| WMT10 (devel) | 15.66 | 15.74 |
| WMT11     | 16.39 | 16.42 |
| WMT12     | 13.81 | 13.85 |
Automatic evaluation (NIST)

- WMT10 (devel): SMT output 5.44, after Deepfix 5.47
- WMT11: SMT output 5.73, after Deepfix 5.74
- WMT12: SMT output 5.26, after Deepfix 5.28
### Automatic evaluation (1-PER)

|           | WMT10 (devel) | WMT11 | WMT12 |
|-----------|---------------|-------|-------|
| SMT output| 41.56         | 42.83 | 39.96 |
| after Deepfix | 41.74         | 42.91 | 40.09 |
Manual evaluation (changed stcs)

- Improvement: 321, 56%
- Degradation: 135, 24%
- Indefinite: 113, 20%
Conclusion

- address valency errors
  - statistical post-editing of SMT
- identify head-argument pairs (DEEP)
  - deep syntactic analysis
- find the best case for the arguments (FIX)
  - statistical valency model
- obtain slight improvement of translation quality
  - indicated by automatic evaluation
  - confirmed by manual evaluation
Future work

- explore existing valency lexicons
- more intricate modelling
  - combine more models
  - machine learning (now thresholds semi-manual, and overfitted to development data)
- further adapt underlying NLP tools (tagger)
- extend to other language pairs
Thank you for your attention

Rudolf Rosa, David Mareček, Aleš Tamchyna
{rosa,marecek,tamchyna}@ufal.mff.cuni.cz

Deepfix: Statistical Post-editing of Statistical Machine Translation Using Deep Syntactic Analysis

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

For this presentation and other information, please visit:

http://ufal.mff.cuni.cz/~rosa/