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
Motivation

- Source text in English:

\textit{EU criticizes not only the Greek government.}
Motivation

- Source text in English:
  \textit{EU criticizes not only the Greek government}

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

- Post-editation by Deepfix:
  
  EU kritizuje nejen řeckou vládu

- Not only the Greek government criticizes EU

- EU criticizes not only the Greek government
DEEPFIX Outline
DEEPFIX Outline

• Problem definition
  ➔ Errors in valency in SMT outputs
DEEPFIX Outline

- Problem definition
  - Errors in valency in SMT outputs
- Related work
DEEPFIX Outline

- Problem definition
  - Errors in valency in SMT outputs
- Related work
- Step 1 (DEEP): Sentence analysis
  - Deep dependency parsing (t-trees)
DEEPFIX Outline

- Problem definition
  - Errors in valency in SMT outputs
- Related work
- Step 1 (DEEP): Sentence analysis
  - Deep dependency parsing (t-trees)
- Step 2 (FIX): Sentence post-editing
  - Statistical model of valency
• Problem definition
  ➔ Errors in valency in SMT outputs
• Related work
• Step 1 (DEEP): Sentence analysis
  ➔ Deep dependency parsing (t-trees)
• Step 2 (FIX): Sentence post-editing
  ➔ Statistical model of valency
• Results
Subject – Object dichotomy

- **English:** position (left/right constituent)
  - EU<sub>left</sub> (Subject) criticizes the government right (Object)

- **Czech:** morphological case (nominative/other)
  - EU<sub>nominative</sub> (Subject) kritizuje vládu<sub>accusative</sub> (Object)
  - vládu<sub>accusative</sub> (Object) kritizuje EU<sub>nominative</sub> (Subject)
  - EU<sub>nominative</sub> (Subject) vládu<sub>accusative</sub> (Object) kritizuje
  - vládu<sub>accusative</sub> (Object) EU<sub>nominative</sub> (Subject) kritizuje
  - (position may mark topic-focus articulation, stress...)
Valency of criticize (kritizovat)

- EU$_{subject}$ *criticizes* the Greek government$_{object}$
- EU$_{nominative}$ *kritizuje* řeckou vládu$_{accusative}$
Valency of \textit{criticize} (\textit{kritizovat})

- $EU_{\text{subject}}$ \textit{criticizes} the \textit{Greek government}_{\text{object}}
- $EU_{\text{nominative}}$ \textit{kritizuje} \textit{řeckou vládu}_{\text{accusative}}
- a valency frame of a verb
  - subject \textit{criticize} object
  - nominative \textit{kritizovat} accusative
Valency of criticize (kritizovat)

- **EU**<sub>subject</sub> *criticizes* the Greek government<sub>object</sub>
- **EU**<sub>nominative</sub> *kritizuje* řeckou vládu<sub>accusative</sub>
- a valency frame of a verb
  - subject *criticize* object (position)
  - nominative *kritizovat* accusative (cases)
Valency of criticize (kritizovat)

- EU\text{subject} \text{criticizes} the Greek government\text{object}
- EU\text{nominative} \text{kritizuje řeckou vládu}\text{accusative}
- a valency frame of a verb
  - subject \text{criticize} object \text{(position)}
  - nominative \text{kritizovat} accusative \text{(cases)}
- decomposition into head-argument pairs
  - (to criticize, government) \sim (kritizovat, vládu)
  - (to criticize, Object) \sim (kritizovat, accusative)
Correction approach: rule-based?

- Rule-based post-editing successful for many types of errors in English-to-Czech translation
  - Morphological agreement, verb tenses, possessive constructions, passive constructions, negation...
  - In → SMT → Depfix → out (watch Monday seminar)
  - Easy to do using Czech positional tagset & analysis of Czech and English to a-trees/t-trees in Treex
Correction approach: rule-based?

- rule-based post-editing successful for many types of errors in English-to-Czech translation
  - morphological agreement, verb tenses, possessive constructions, passive constructions, negation...
  - in → SMT → Depfix → out (watch Monday seminar)
  - easy to do using Czech positional tagset & analysis of Czech and English to a-trees/t-trees in Treex
- hard to fully cover valency by a set of rules
  - we need it parallelly for English and Czech
  - possible future work: use existing valency lexicons
Correlation approach: statistical?

- Statistical machine translation (SMT) works well
- Statistical post-editing of rule-based machine translation (RBMT) outputs works well
  - in → RBMT → SMT → out (Simard et al., 2007)
- Statistical post-editing of SMT outputs
  - in → SMT → SMT → out
    - works for English-to-Turkish (Oflazer et al., 2007)
    - works for French-to-English (Béchara et al., 2011)
Correction approach: statistical?

- Statistical machine translation (SMT) works well
- Statistical post-editing of rule-based machine translation (RBMT) outputs works well
  - in → RBMT → SMT → out (Simard et al., 2007)
- Statistical post-editing of SMT outputs
  - in → SMT → SMT → out
  - works for English-to-Turkish (Oflazer et al., 2007)
  - works for French-to-English (Béchara et al., 2011)
  - does not work for English-to-Czech (Tamchyna)
Correction approach: combine!

- capturing valency by rules – not good
  - let's use statistics for that!
- simple statistical post-editing of SMT – not good
  - let's get some inspiration from the linguistically motivated rule-based approaches!
- Step 1: analyze the sentences in Treex
  - linguistically motivated, combines rules and statistics
- Step 2: correct valency with a statistical model
  - simple statistics, but operating on the t-layer
Sentence analysis (DEEP)

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

EU criticizes the Greek government

EU criticizes the Greek government

EU kritizuje řecká vláda

EU kritizovat řecká vláda
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
(head, arg) pair identification

EU criticizes the Greek government

EU kritizuje řecká vláda

EU N, subject

government N, object

Greek A, attribute

kritizovat V, predicate

EU N, nominative

vláda N, nominative

řecká A, attribute
Valency models (FIX)

- $P(\text{arg}_{\text{case}} \mid \text{head}_{\text{lemma}}, \text{English}_{\text{arg}}_{\text{function}})$
- $P(\text{arg}_{\text{case}} \mid \text{head}_{\text{lemma}}, \text{English}_{\text{arg}}_{\text{function}}, \text{arg}_{\text{lemma}})$
- 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

- $P(\text{nominative} \mid kritizovat, \text{object}) = 0.03$
- $P(\text{accusative} \mid kritizovat, \text{object}) = 0.80$
- 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 kritizuje nejen řecká nominative vláda nominative*

  - Not only *the Greek government* criticizes *EU*

  - Valency model correction:

  *EU kritizuje nejen řecká nominative vládu accusative*
Sentence correction

- Statitical machine translation output:

$$EU \text{ kritizuje nejen } \text{řecká}^{\text{nominative}} \text{ vláda}^{\text{nominative}}$$

  - Not only **the Greek government** criticizes **EU**

- Valency model correction:

$$EU \text{ kritizuje nejen } \text{řecká}^{\text{nominative}} \text{ vládu}^{\text{accusative}}$$

- Agreement enforcement:

$$EU \text{ kritizuje nejen } \text{řeckou}^{\text{accusative}} \text{ vládu}^{\text{accusative}}$$

  - **EU criticizes not only the Greek government**
Some interesting details

- the model actually works on formemes
  - functions (EN), cases (CS), **prepositions** (EN, CS)
  - in: *The government spends on the middle schools.*
  - SMT: *Vládá utrácí střední školy.*
    - *(spend, on+X) → (utrácet, 4)  P = 0.07*
    - *The government destroys the middle schools.*
  - out: *Vládá utrácí za střední školy.*
    - *(spend, on+X) → (utrácet, za+4)  P = 0.89*
    - *The government spends on the middle schools.*
Some interesting details

- the model actually works on formemes
  - functions (EN), cases (CS), prepositions (EN, CS)
  - in: *The government spends on the middle schools.*
  - SMT: *Vládá utrácí střední školy.*

  - (spend, on+X) → (utrácet, 4) \( P = 0.07 \)

  - *The government destroys the middle schools.*

  - out: *Vládá utrácí za střední školy.*

  - (spend, on+X) → (utrácet, za+4) \( P = 0.89 \)

  - *The government spends on the middle schools.*

- we model both verb valency and noun valency
Automatic evaluation (BLEU)

- WMT10 (devel): 15.66, 15.74
- WMT11: 16.39, 16.42
- WMT12: 13.81, 13.85

SMT output vs. after Deepfix
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
Possible future work

- 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
- explore existing valency lexicons
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/