Depfix:

Automatic post-editing of phrase-based machine translation outputs

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

Monday seminar, 14th October 2013
Outline

- translation of negation (and its correction)
  - motivation
  - the fixing pipeline
- analysis and corrections
  - m-layer (lemmas, tags, word-alignment)
  - a-layer (dependency trees, analytical functions)
  - t-layer (“tecto-trees”, formemes, grammatemes)
- evaluation
- parsing of SMT outputs (MSTperl parser)
Motivation: Translation of negation
Motivation: Errors in negation

- These are not actually errors.
Motivation: Errors in negation

- These are not actually errors.
  - Moses: **Jsou** to vlastně chyby.
  - Gloss: These are actually errors.
Motivation: Errors in negation

- These **are not** actually errors.
  - Moses: **Jsou** to vlastně chyby.
  - Gloss: **These are** actually errors.
  - Ref.: **Nejsou** to vlastně chyby.
Motivation: Errors in negation

- These are not actually errors.
  - Moses: Jsou to vlastně chyby.
  - Gloss: These are actually errors.
  - Ref.: Nejsou to vlastně chyby.
- I would not cheat on you.
Motivation: Errors in negation

- These are not actually errors.
  - Moses: J*so*u to vlastně chyby.
  - Gloss: These are actually errors.
  - Ref.: Nejsou to vlastně chyby.

- I would not cheat on you.
  - Moses: Já bych tě podváděl.
  - Gloss: I would cheat on you.
  - Ref.: Já bych tě nepodváděl.
Motivation: Errors in negation

- These are not actually errors.
  - Moses: Jsou to vlastně chyby.
  - Gloss: These are actually errors.
  - Ref.: Nejsou to vlastně chyby.
- I would not cheat on you.
  - Moses: Já bych tě podváděl.
  - Gloss: I would cheat on you.
  - Ref.: Já bych tě nepodváděl.
- Some phenomena hard to get right with PBSMT
Simple to fix?

- there is a negation in the source
  - These are not actually errors
Simple to fix?

- there is a negation in the source
  - These are not actually errors
- there is no negation in the target
  - Jsou to vlastně chyby
Simple to fix?

- There is a negation in the source
  - These are **not** actually errors
- There is no negation in the target
  - Jsou to vlastně chyby
Simple to fix?

- there is a negation in the source
  - These are **not** actually errors
- there is no negation in the target
  - *Jsou to vlastně chyby*
- add the negative prefix (ne-) into the target
  - *Nejsou to vlastně chyby*
Simple to fix?

- there is a negation in the source
  - These are not actually errors
- there is no negation in the target
  - Jsou to vlastně chyby
- add the negative prefix (ne-) into the target
  - Nejsou to vlastně chyby
- such a simple approach might be sufficient
- but usually useful to use some more NLP tools
Part-of-speech tagger

- run a POS tagger on the target sentence
  - Jsou to vlastně chyby
  - verb pronoun adverb noun

- a good heuristic: negate the (finite) verb!
  - Nejsou to vlastně chyby
  - verb pronoun adverb noun

- fine-grained tags may even mark the negation
  - jsou VB-P---3P-AA---
  - nejsou VB-P---3P-NA---
Dependency parser

- parse both source and target
- project negation through word alignment
Dependency parser

- parse both source and target
- project negation through word alignment
Deep syntactic analysis

- auxiliary nodes collapsed into values of attributes on parent nodes

- abstract from various ways of expressing negation (not, no, un-, in-,...)
  - all marked by \texttt{neg=1} on the lexical node
Morphological generator

- \texttt{form = generate(word, morphological features)}
Morphological generator

- `form = generate(lemma, tag)`
Morphological generator

- \texttt{form = generate(lemma, tag)}
- \textbf{instead of}: \texttt{new\_form = 'ne' + form}
  - \texttt{'nejsou' = 'ne' + 'jsou'}
Morphological generator

- `form = generate(lemma, tag)`
- **instead of:** `new_form = 'ne' + form`
  - `'nejsou' = 'ne' + 'jsou'`
- **use the more sophisticated:** `new_form = generate(lemma(form), negate(tag))`
  - `'nejsou' = generate(lemma('jsou'), negate('VB-P---3P-AA---'))`
Morphological generator

- form = generate(lemma, tag)
- instead of: new_form = 'ne' + form
  - 'nejsou' = 'ne' + 'jsou'
- use the more sophisticated: new_form = generate(lemma(form), negate(tag))
  - 'nejsou' = generate(lemma('jsou'), negate('VB-P---3P-AA---'))
  - 'nejsou' = generate('být', 'VB-P---3P-NA---')
Depfix pipeline

- input
  - English sentence
  - its Czech translation, provided by an SMT system
- m-layer analysis and fixes (tagger, word-aligner)
- a-layer analysis and fixes (dependency parser)
- t-layer analysis and fixes (t-tree analyzer)
- output (morphological generator)
  - the Czech translation, corrected
Outline

- translation of negation (and its correction)
  - motivation
  - the fixing pipeline
- analysis and corrections
  - m-layer (lemmas, tags, word-alignment)
  - a-layer (dependency trees, analytical functions)
  - t-layer (“tecto-trees”, formemes, grammatemes)
- evaluation
- parsing of SMT outputs (MSTperl parser)
Outline

✔ translation of negation (and its correction)
  ✔ motivation
  ✔ the fixing pipeline
→ analysis and corrections
  → m-layer (lemmas, tags, word-alignment)
    ▪ a-layer (dependency trees, analytical functions)
    ▪ t-layer (“tecto-trees”, formemes, grammatemes)
  ▪ evaluation
  ▪ parsing of SMT outputs (MSTperl parser)
M-layer

- analysis: lemmas, tags, word-alignment
  - adding missing alignment links by string similarity
- corrections:
  - tokenization projection
  - morphological number projection
  - source-aware truecasing
  - vocalisation of prepositions
<table>
<thead>
<tr>
<th>Source</th>
<th>Michèle Alliot-Marie had sent a communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses</td>
<td>Michèle Alliot - Marie poslal sdělení</td>
</tr>
<tr>
<td>Gloss</td>
<td>Michèle Alliot - Marie sent_{masc} a communication</td>
</tr>
<tr>
<td>Depfix</td>
<td>Michèle Alliot-Marie poslala sdělení</td>
</tr>
<tr>
<td>Gloss</td>
<td>Michèle Alliot-Marie sent_{fem} a communication</td>
</tr>
<tr>
<td>Source</td>
<td>the director of the best hotel in <strong>Pec</strong>, Karel <strong>Rada</strong>.</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------</td>
</tr>
<tr>
<td>Moses</td>
<td>ředitel nejlepší hotel v <strong>peci</strong>, Karel <strong>rada</strong>.</td>
</tr>
<tr>
<td>Gloss</td>
<td>the director of the best hotel in the <strong>oven</strong>, Karel <strong>advice</strong>.</td>
</tr>
<tr>
<td>Depfix</td>
<td>ředitel nejlepšího hotelu v <strong>Peci</strong>, Karel <strong>Rada</strong>.</td>
</tr>
<tr>
<td>Gloss</td>
<td>the director of the best hotel in <strong>Pec</strong>$<em>{\text{town}}$, Karel **Rada$</em>{\text{surname}}$.**</td>
</tr>
</tbody>
</table>
Vocalisation of prepositions

<table>
<thead>
<tr>
<th>Source</th>
<th>The work being done by experts from three institutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses</td>
<td>Práce odborníků z tří institucí</td>
</tr>
<tr>
<td>Gloss</td>
<td>Work by experts <strong>from</strong> three institutions</td>
</tr>
<tr>
<td>Depfix</td>
<td>Práce odborníků <strong>ze</strong> tří institucí</td>
</tr>
<tr>
<td>Gloss</td>
<td>Work by experts <strong>from</strong> three institutions</td>
</tr>
</tbody>
</table>
Outline

✔ translation of negation (and its correction)
  ✔ motivation
  ✔ the fixing pipeline

➔ analysis and corrections
  ✔ m-layer (lemmas, tags, word-alignment)
    ➔ a-layer (dependency trees, analytical functions)
      ▪ t-layer (“tecto-trees”, formemes, grammatemes)

▪ evaluation

▪ parsing of SMT outputs (MSTperl parser)
A-layer

- analysis: dependency trees, analytical functions
  - for Czech: MSTperl (adapted MST parser)
  - + fixing prepositions without children, auxiliary verbs with children, AuxT/AuxR...
- morphological agreement fixes:
  - preposition-noun, noun-adjective, subject-predicate...
- fixes of transfer of meaning to morphology
  - possessives, passives, subject...
A-layer motivation

- **Source:**
  - *All the winners received a diploma.*

- **Moses:**
  - *Všem výhercům obdržel diplom.*
  - *To all the winners he received a diploma.*

- **Depfix:**
  - *Všichni výherci obdrželi diplom.*
  - *All the winners received a diploma.*
Všem výhercům obdržel diplom.
Transfer of meaning: subject

a-tree zone=en
received Pred VBD
winners Sb NNS
diploma Obj NN
All Atr PDT the AuxA DT a AuxA DT
Všem Atr PLXP3

a-tree zone=cs
obdržel Pred VpYSXRA
výhercům Obj NNMP3
diplom Obj NNIS1
Transfer of meaning: subject

a-tree
zone=en

received
Pred
VBD

winners
Sb
NNS

diploma
Obj
NN

AuxA
DT

AuxK

AuxK

Všem
Atr
PLXP3

All
Atr
PDT

the
AUXA
DT

a
AUXA
DT

výhercům
Obj
NN

obdržel
Pred
VpYSXRA

diplom
Obj
NN

zone=cs
Rudolf Rosa – Depfix: Automatic post-editing of phrase-based machine translation outputs
Všem výherci obdržel diplom.
Noun-adjective agreement
Agreement: gender, case (number)

- The winners received diplomas.
- All Atr PDT
- the AuxA DT
- a AuxA DT
- Všichni Atr PLMP1
- diplom Obj NNIS1
- obdržel Pred VpYSXRA
- výherci Sb NNMP1
- obdržel Pred VpYSXRA
- All Atr PDT
- the AuxA DT
- a AuxA DT
- Všichni Atr PLMP1
- diplom Obj NNIS1

Rudolf Rosa, Depfix: Automatic post-editing of phrase-based machine translation outputs
Všichni výherci obdržel diplom.
Subject-predicate agreement

Rudolf Rosa – Depfix: Automatic post-editing of phrase-based machine translation outputs
Agreement: gender, num (person)
Všichni výherci obdrželi diplom.
## Preposition-noun agreement

<table>
<thead>
<tr>
<th>Source</th>
<th>It is a story about sport, race relations, and Nelson Mandela.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses</td>
<td>Je to příběh o sportu, rasových vztahů, a Nelson Mandela.</td>
</tr>
<tr>
<td>Gloss</td>
<td>It is a story about sport, race relations, and Nelson Mandela.</td>
</tr>
<tr>
<td>Depfix</td>
<td>Je to příběh o sportu, rasových vztazích, a Nelson Mandelovi.</td>
</tr>
<tr>
<td>Gloss</td>
<td>It is a story about sport, race relations, and Nelson Mandela.</td>
</tr>
<tr>
<td>Source</td>
<td>this half-hearted increase will bear the same fruit</td>
</tr>
<tr>
<td>--------</td>
<td>---------------------------------------------------</td>
</tr>
<tr>
<td>Moses</td>
<td>tato polovičatá nárůst bude nést stejné ovoce</td>
</tr>
<tr>
<td>Gloss</td>
<td>this\textsubscript{fem} half-hearted\textsubscript{fem} increase\textsubscript{masc} will bear the same fruit</td>
</tr>
<tr>
<td>Depfix</td>
<td>tento polovičatý nárůst bude nést stejné ovoce</td>
</tr>
<tr>
<td>Gloss</td>
<td>this\textsubscript{masc} half-hearted\textsubscript{masc} increase\textsubscript{masc} will bear the same fruit</td>
</tr>
<tr>
<td>Source</td>
<td>unsustainable deficit level of public finances.</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>Moses</td>
<td>neudržitelná úroveň schodku veřejné finance.</td>
</tr>
<tr>
<td>Gloss</td>
<td>unsustainable deficit level public finances.</td>
</tr>
<tr>
<td>Depfix</td>
<td>neudržitelná úroveň schodku veřejných financí.</td>
</tr>
<tr>
<td>Gloss</td>
<td>unsustainable deficit level of public finances.</td>
</tr>
<tr>
<td>Source</td>
<td>Janota's possible <strong>continuation</strong> in office will be the topic of Friday's meeting.</td>
</tr>
<tr>
<td>--------</td>
<td>-------------------------------------------------------------------</td>
</tr>
<tr>
<td>Moses</td>
<td>Janota je možné <strong>pokračování</strong> ve funkci bude tématem páteční schůze.</td>
</tr>
<tr>
<td>Gloss</td>
<td>Janota is possible <strong>continuation</strong> in office will be the topic of Friday's meeting.</td>
</tr>
<tr>
<td>Depfix</td>
<td>možné <strong>pokračování Janoty</strong> ve funkci bude tématem páteční schůze.</td>
</tr>
<tr>
<td>Gloss</td>
<td>possible <strong>continuation of Janota</strong> in office will be the topic of Friday's meeting.</td>
</tr>
</tbody>
</table>
## Translation of Subject

<table>
<thead>
<tr>
<th>Source</th>
<th>At a time when Swiss voters(_{\text{subj}}) have called for a ban on the construction of minarets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses</td>
<td>V době, kdy švýcarské voliče(_{\text{acc}}) vyzvali k zákazu výstavby minaretů</td>
</tr>
<tr>
<td>Gloss</td>
<td>At a time, when Swiss voters were called for a ban on the construction of minarets</td>
</tr>
<tr>
<td>Depfix</td>
<td>V době, kdy švýcarští voliči(_{\text{nom}}) vyzvali k zákazu výstavby minaretů</td>
</tr>
<tr>
<td>Gloss</td>
<td>At a time, when Swiss voters called for a ban on the construction of minarets</td>
</tr>
</tbody>
</table>
Outline

✔ translation of negation (and its correction)
  ✔ motivation
  ✔ the fixing pipeline

➔ analysis and corrections
  ✔ m-layer (lemmas, tags, word-alignment)
  ✔ a-layer (dependency trees, analytical functions)
  ➔ t-layer (“tecto-trees”, formemes, grammatemes)

▪ evaluation

▪ parsing of SMT outputs (MSTperl parser)
T-layer

- analysis: t-trees, formemes, grammatemes
  - + systematic analysis of English verb tenses
- rule-based fixes:
  - negation
  - verb tenses
  - subject pronoun dropping
- statistical fixes:
  - valency
## Translation of verb tenses

<table>
<thead>
<tr>
<th>Source</th>
<th>This <strong>will bring</strong> problems for whoever is in office</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses</td>
<td>To <strong>přináší</strong> problémy pro každého, kdo je v kanceláři</td>
</tr>
<tr>
<td>Gloss</td>
<td>This <strong>brings</strong> problems for anyone who is in office</td>
</tr>
<tr>
<td>Depfix</td>
<td>To <strong>bude přinášet</strong> problémy pro každého, kdo je v kanceláři</td>
</tr>
<tr>
<td>Gloss</td>
<td>This <strong>will bring</strong> problems for anyone who is in office</td>
</tr>
</tbody>
</table>
### Subject pronoun dropping

<table>
<thead>
<tr>
<th>Source</th>
<th>I don't blame them.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses</td>
<td>Já se jim nedivím.</td>
</tr>
<tr>
<td>Gloss</td>
<td>I myself them don't-blame&lt;sub&gt;1st person sg&lt;/sub&gt;.</td>
</tr>
<tr>
<td>Depfix</td>
<td>Nedivím se jim.</td>
</tr>
<tr>
<td>Gloss</td>
<td>Don't-blame&lt;sub&gt;1st person sg&lt;/sub&gt; myself them.</td>
</tr>
</tbody>
</table>
Correction of valency errors

- 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*\textsuperscript{nom}

  - *Not only the Greek government criticizes EU*

- Post-editation by Deepfix:
  
  *EU kritizuje nejen řeckou vládu*\textsuperscript{acc}

  - *EU criticizes not only the Greek government*
Valency of criticize (kritizovat)

- **EU**_{subject} **criticizes** the Greek government_{object}
- **EU**_{nominative} **kritizuje** řeckou vládu_{accusative}
- a valency frame of a verb
  - subject **criticize** object (position)
  - nominative **kritizovat** accusative (cases)
- decomposition into head-argument pairs
  - (to criticize, government) ~ (kritizovat, vládu)
  - (to criticize, Object) ~ (kritizovat, accusative)
EU criticizes the Greek government

EU kritizuje řecká vláda

Deep syntactic dependency trees
Deep syntactic dependency trees

EU criticizes the Greek government

EU kritizuje řecká vláda

EU criticize government

EU kritizovat vláda

Greek řecká

EU criticizes the Greek government

EU kritizuje řecká vláda

EU criticize government

EU kritizovat vláda

Greek řecká
Deep syntactic dependency trees

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
řecká
A, attribute
(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

řecká A, attribute
Valency models (FIX)

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

- \( P(\text{nominative} \mid \text{kritizovat}, \text{object}) = 0.03 \)
- \( P(\text{accusative} \mid \text{kritizovat}, \text{object}) = 0.80 \)
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
Statitical machine translation output:

*EU kritizuje nejen řecká vláda*

- Not only *the Greek government* criticizes EU

Valency model correction:

*EU kritizuje nejen řecká vládu*

- Agreement enforcement:

*EU kritizuje nejen řeckou vládu*

- 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
Outline

✔ translation of negation (and its correction)
  ✔ motivation
  ✔ the fixing pipeline
✔ analysis and corrections
  ✔ m-layer (lemmas, tags, word-alignment)
  ✔ a-layer (dependency trees, analytical functions)
  ✔ t-layer (“tecto-trees”, formemes, grammatemes)

➔ evaluation

▪ parsing of SMT outputs (MSTperl parser)
Automatic evaluation (BLEU)

<table>
<thead>
<tr>
<th>Year</th>
<th>SMT output</th>
<th>After Depfix</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT10 (devel)</td>
<td>15.66</td>
<td>16.08</td>
</tr>
<tr>
<td>WMT11</td>
<td>16.39</td>
<td>16.61</td>
</tr>
<tr>
<td>WMT12</td>
<td>13.81</td>
<td>13.85</td>
</tr>
<tr>
<td>WMT13</td>
<td>20.08</td>
<td>20.02</td>
</tr>
</tbody>
</table>
Manual evaluation

- changed sentences
  - Improvement: 430 (58%)
  - Degradation: 152 (21%)
  - Indefinite: 157 (21%)

- all sentences
  - No change: 611 (45%)
  - Improvement: 430 (32%)
  - Degradation: 152 (11%)
  - Indefinite: 157 (12%)

Rudolf Rosa – Depfix: Automatic post-editing of phrase-based machine translation outputs
Manual evaluation (WMT13)

- TectoMT: 0.49
- TectoMT+Moses: 0.56
- TectoMT+Moses+Depfix: 0.54
- Google Translate: 0.53

Scientists: 0.64, 0.66, 0.62

Mechanical Turkers: 0.46, 0.56, 0.53
Jak překladač z Matfyzu porazil Google

(...)

Pomohl odstraňovač chyb

(...)

Aby byl výsledek ještě lepší, prošel výsledný text ještě automatickou korekturou pomocí českého systému Depfix, odstraňovače chyb, jenž opravil například špatně přeložené negativní věty a pády.

http://tech.ihned.cz/hnfuture/c1-60978500-prekladac-google-maffyz-system
Precision of rules

$$precision = \frac{|improved|}{|improved| + |worsened|}$$
Impact of rules

\[
\text{impact} = \frac{|\text{modified}|}{|\text{evaluated}|}
\]

75% precision 60% precision 50% precision

5% impact
Outline

✔ translation of negation (and its correction)
  ✔ motivation
  ✔ the fixing pipeline
✔ analysis and corrections
  ✔ m-layer (lemmas, tags, word-alignment)
  ✔ a-layer (dependency trees, analytical functions)
  ✔ t-layer (“tecto-trees”, formemes, grammatemes)
✔ evaluation

➔ parsing of SMT outputs (MSTperl parser)
Parsing of SMT Outputs

- can be useful in many applications
  - automatic classification of translation errors
  - automatic correction of translation errors (Depfix)
  - multilingual question answering...

✔ we have the source sentence available
  - Can we use it to help parsing?

❌ SMT outputs noisy (errors in fluency, grammar...)
  - parsers trained on gold standard treebanks
  - Can we adapt parser to noisy sentences?
MST parser

- Maximum Spanning Tree parser
- McDonald, Crammer, Pereira (2005)
  - Online large-margin training of dependency parsers
- McDonald, Pereira, Ribarov, Hajič (2006)
  - Non-projective dependency parsing using spanning tree algorithms
(1) Words and Tags

words = nodes

# root

relaxes VBZ

Rudolph NNP

abroad RB
(2) (Nearly) Complete Graph

all possible edges = directed edges

Rudolph
NNP

relaxes
VBZ

abroad
RB

# root
(3) Assign Edge Weights

edge weight = sum of edge features weights

Margin Infused Relaxed Algorithm (MIRA)
(4) Maximum Spanning Tree

non-projective trees: Chu-Liu-Edmonds algorithm

(projective trees: Eisner algorithm)
dependency tree = maximum spanning tree

Rudolph

relaxes

abroad

root

VBZ

dependency tree =

maximum spanning tree
labels assigned by a second stage labeler
Tool::Parser::MSTperl

- reimplementation of MST Parser in Perl
  - Treex::Block::W2A::CS::ParseMSTperl
  - (so far only) first-order, non-projective
- adapted for SMT outputs parsing
  - worsening the training data (David Mareček)
  - adding parallel information
  - manually boosting feature weights
  - exploiting large-scale data
Parser Training Data

- Prague Czech-English Dependency Treebank
  - parallel treebank
  - 50k sentences, 1.2M words
  - morphological tags, surface syntax, deep syntax
  - word alignment
Worsening the Treebank

- treebank used for training contains correct sentences
- SMT output is noisy
  - grammatical errors
  - incorrect word order
  - missing/superfluous words
  - …
- let's introduce similar errors into the treebank!
  - so far, we have only tried inflection errors
translate **English** side of PCEDT to **Czech**
- by an SMT system (we used Moses)

now we have e.g.:
- **Gold English**
  - Rudolph's car is black.
- **Gold Czech**
  - Rudolfovo auto je černé.
- **SMT Czech**
  - Rudolfova auto je černý.
Worsen (2): Align SMT to Gold Czech

- align SMT Czech to Gold Czech
- Monolingual Greedy Aligner (Martin Popel)
  - alignment link score = linear combination of:
    - similarity of word forms (or lemmas)
    - similarity of morphological tags (fine-grained)
    - similarity of positions in the sentence
    - indication whether preceding/following words aligned
  - repeat: align best scoring pair until below threshold
  - no training: weights and threshold set manually
Worsen (3): Create Error Model

- for each tag:
  - estimate probabilities of SMT system using an incorrect tag instead of the correct tag (Maximum Likelihood Estimate)
- Czech tagset: fine-grained morphological tags
  - part-of-speech, gender, number, case, person, tense, voice...
  - 1500 different tags in training data
Adjective, Masculine, Plural, Instrumental case (AAMP7), e.g. *lingvistickými* (linguistic)

- 0.2 Adjective, Masculine, Singular, Nominative case
  - e.g. *lingvistický*

- 0.1 Adjective, Masculine, Plural, Nominative case
  - e.g. *lingvističtí*

- 0.1 Adjective, Neuter, Singular, Accusative case
  - e.g. *lingvistiké*

... altogether 2000 such change rules
Worsen (4): Apply Error Model

- take Gold Czech
- for each word:
  - assign a new tag randomly sampled according to Tag Error Model
  - generate a new word form
    - rule-based generator, generates even unseen forms
    - new_form = generate_form(lemma, tag) || old_form
- → get Worsened Czech
- use resulting Gold English-Worsened Czech parallel treebank to train the parser
Parallel Features

- word alignment (using GIZA++)
- additional features (if aligned node exists):
  - aligned tag (NNS, VBD...)
  - aligned dependency label (Subject, Attribute...)
  - aligned edge existence (0/1)
Parallel Features Example

Rudolf
NN N M S 1

Relaxuje
VB S 3

Rudolph
NNP

Relaxes
VBZ

Abroad
RB

Zahraničí
NN N S 6

V
RR 6

Pred

Subj

AuxP

Adv

Root

Pred
Manually boosting feature weights

- **aligned edge existence** is the key feature here
- observation: since the worsening is probably too mild, its weight is too low
  - edge exists: -0.57
  - edge does not exist: -0.83
  - missing aligned node(s): -0.67
Manually boosting feature weights

- **aligned edge existence** is the key feature here
- observation: since the worsening is probably too mild, its weight is too low
  - edge exists: -0.57
  - edge does not exist: -0.83
  - missing aligned node(s): -0.67
- experiment: manually increase its weight
  - edge exists: -0.25
Manually boosting feature weights

- **aligned edge existence** is the key feature here
- observation: since the worsening is probably too mild, its weight is too low
  - edge exists: **-0.57**
  - edge does not exist: **-0.83**
  - missing aligned node(s): **-0.67**
- experiment: manually increase its weight
  - edge exists: **-0.25**
- success – manual changing of weights feasible
Exploiting large-scale data

- exploiting large-scale parsed data (CzEng) to provide additional lexical features
- lexical features are important for the parser
- CzEng has 10 times more word types (lemmas) than PCEDT (400k vs. 40k)
- training the parser on whole CzEng infeasible
- new feature: pointwise mutual information

\[ PMI'(parent, child) = \log \frac{\text{count}([parent, child])}{\text{count}([parent, *]) \cdot \text{count}([*, child])} \]
Direct Evaluation: by Inspection

- manual inspection of several parse trees
  - comparing baseline and adapted parser outputs
- examples of improvements:
  - subject identification even if not in nominative case
  - adjective-noun dependence identification even if agreement violated (gender, number, case)
- hard to do reliably
  - trying to find a correct parse tree for an (often) incorrect sentence – not well defined
Indirect Evaluation: in Depfix

- improvements and deteriorations in comparison to Depfix employing a baseline parser:
  - original McDonald's MST Parser
  - our baseline setup, without adaptations
Outline – all done!

✔ translation of negation (and its correction)
  ✔ motivation
  ✔ the fixing pipeline
✔ analysis and corrections
  ✔ m-layer (lemmas, tags, word-alignment)
  ✔ a-layer (dependency trees, analytical functions)
  ✔ t-layer (“tecto-trees”, formemes, grammatemes)
✔ evaluation
✔ parsing of SMT outputs (MSTperl parser)
Thank you for your attention

Rudolf Rosa
rosa@ufal.mff.cuni.cz

Depfix:
Automatic post-editing
of phrase-based machine translation outputs

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/