Machine Learning for Deep-syntactic MT

Martin Popel

ÚFAL (Institute of Formal and Applied Linguistics)
Charles University in Prague

September 11, 2015
Seminar on the 35th Anniversary of the Cooperation between
Charles University in Prague and Hamburg University
1 Intro
   - TectoMT schema
   - Isomorphic transfer

2 MT as labeling

3 TectoMT over years
   - 2008 baseline transfer
   - 2009 HMTM
   - 2010 MaxEnt
   - 2014 VowpalWabbit

4 Future plans
TectoMT: analysis, transfer, synthesis

**ANALYSIS**
- Tectogrammatical layer
  - fill formems
  - grammatemes
  - build t-tree
  - mark edges to contract
  - analytical layer
    - analytical functions
    - dependency parser (MST)
  - morphological layer
    - tagger (MorphoDiTa)
    - lemmatization
    - tokenization
    - segmentation

**TRANSFER**
- t-layer
  - query dictionary
  - use HMTM
  - fill morphological categories
  - impose agreement
  - add functional words

**SYNTHESIS**
- a-layer
  - generate wordforms
- m-layer
- w-layer
  - concatenate

**Blocks**
- Source language (English)
- Target language (Czech)

**Types of blocks**
- Rule based
- Statistical
TectoMT: isomorphic transfer (1-1 node mapping)

Agátha přišla ta kniha zajímavá.

Agátha found that book interesting.

(adj:compl means predicative adjective)
Representation of t-layer

**lemma and formeme as two attributes**

- **find**
  - **v:fin**
  - **Agatha**
    - **n:subj**
  - **book**
    - **n:obj**
  - **this**
    - **adj:attr**
  - **interesting**
    - **adj:compl**
lemma and formeme as interleaved “sub-nodes”
Representation of t-layer

lemma and formeme as interleaved “sub-nodes”

v:fin

find

n:subj Agatha

n:obj book

adj:attr this

adj:compl interesting

přijít

n:3

n:1

adj:1
Representation of t-layer

**lemma and formeme** as interleaved “sub-nodes”

```
    v:fin
     |
  find
 /|
n:subj  n:obj  adj:compl
Agatha book interesting
    |
adj:attr
    |
this
```

**grammatemes:**
- translated in postprocessing (current approach)
- as subnodes (leaves, children of lemmas)
- encoded within lemma, but only if grammateme changed
Handling non-isomorphic transfer

- preprocessing or postprocessing within transfer (current approach)
- natively in the main transfer algorithm
- convert training data to isomorphic trees [not tried yet]
  - n-1 alignment: add special [delete_node] label to the target side
  - 1-n alignment: encode added nodes (L+F) into the “main” lemma
  - encode topology change: as_child, as_sibling, as_parent
TectoMT transfer over years

<table>
<thead>
<tr>
<th>year</th>
<th>BLEUdiff</th>
<th>method</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td></td>
<td>initial baseline</td>
</tr>
<tr>
<td>2009</td>
<td>+1.5</td>
<td>HMTM (TreeViterbi, TreeLM)</td>
</tr>
<tr>
<td>2010</td>
<td>+0.8</td>
<td>HMTM + MaxEnt</td>
</tr>
<tr>
<td>2012</td>
<td>-2.2</td>
<td>TectoMoses</td>
</tr>
<tr>
<td>2012</td>
<td>NA</td>
<td>Gibbs sampling treelets</td>
</tr>
<tr>
<td>2013</td>
<td>-3.0</td>
<td>Easy-first treelets</td>
</tr>
<tr>
<td>2013</td>
<td>-2.0</td>
<td>Interpol treelets</td>
</tr>
<tr>
<td>2014</td>
<td>+0.1</td>
<td>VowpalWabbit</td>
</tr>
</tbody>
</table>
## TectoMT transfer over years

<table>
<thead>
<tr>
<th>year</th>
<th>BLEUdiff</th>
<th>method</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td></td>
<td>initial baseline</td>
</tr>
<tr>
<td>2009</td>
<td>+1.5</td>
<td>HMTM (TreeViterbi, TreeLM)</td>
</tr>
<tr>
<td>2010</td>
<td>+0.8</td>
<td>HMTM + MaxEnt</td>
</tr>
<tr>
<td>2012</td>
<td>-2.2</td>
<td>TectoMoses</td>
</tr>
<tr>
<td>2012</td>
<td>NA</td>
<td>Gibbs sampling treelets</td>
</tr>
<tr>
<td>2013</td>
<td>-3.0</td>
<td>Easy-first treelets</td>
</tr>
<tr>
<td>2013</td>
<td>-2.0</td>
<td>Interpol treelets</td>
</tr>
<tr>
<td>2014</td>
<td>+0.1</td>
<td>VowpalWabbit</td>
</tr>
<tr>
<td></td>
<td>+0.9</td>
<td>other improvements in 2010–2014</td>
</tr>
<tr>
<td>year</td>
<td>BLEUdiff</td>
<td>method</td>
</tr>
<tr>
<td>------</td>
<td>----------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td>initial baseline</td>
</tr>
<tr>
<td>2009</td>
<td>+1.5</td>
<td>HMTM (TreeViterbi, TreeLM)</td>
</tr>
<tr>
<td>2010</td>
<td>+0.8</td>
<td>HMTM + MaxEnt</td>
</tr>
<tr>
<td>2012</td>
<td>-2.2</td>
<td>TectoMoses</td>
</tr>
<tr>
<td>2012</td>
<td>NA</td>
<td>Gibbs sampling treelets</td>
</tr>
<tr>
<td>2013</td>
<td>-3.0</td>
<td>Easy-first treelets</td>
</tr>
<tr>
<td>2013</td>
<td>-2.0</td>
<td>Interpol treelets</td>
</tr>
<tr>
<td>2014</td>
<td>+0.1</td>
<td>VowpalWabbit</td>
</tr>
<tr>
<td></td>
<td>+0.9</td>
<td>other improvements in 2010–2014</td>
</tr>
<tr>
<td>2015</td>
<td>+8.6</td>
<td>QTLeap en→cs in two months</td>
</tr>
</tbody>
</table>
2008: baseline TectoMT transfer

- “static” translation model $P(target|source) = \frac{\#(source, target)}{\#(source)}$
- first translate formemes, then lemmas
- use only the top variant

WMT 2009 en→cs results

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>human score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses (CUNI)</td>
<td>14.2</td>
<td>61</td>
</tr>
<tr>
<td>Google</td>
<td>13.6</td>
<td>66</td>
</tr>
<tr>
<td>Moses (UEdin)</td>
<td>13.5</td>
<td>53</td>
</tr>
<tr>
<td>Eurotran XP</td>
<td>9.5</td>
<td>67</td>
</tr>
<tr>
<td>PC Translator</td>
<td>9.4</td>
<td>67</td>
</tr>
<tr>
<td>TectoMT</td>
<td>7.3</td>
<td>48</td>
</tr>
</tbody>
</table>
2009: Hidden Markov Tree Model (HMTM)

- still using “static” translation models, but also
- TreeLM (target lemma-formeme and parent-child compatibility)
- best labeling is found via HMTM (Tree Viterbi)

Source sentence: 
Strojový překlad by měl být snadný.

Target sentence: 
Machine translation should be easy.

\[ P_E(\text{source | target}) \ldots \text{emission probabilities} \ldots \text{translation model} \]
\[ P_T(\text{dependent | governing}) \ldots \text{transition probabilities} \ldots \text{target-language tree model} \]
2010: Maximum Entropy translation model

- still using HMTM (and generative TreeLM),
- but the “static” model $P(\text{lemma} | \text{src\_lemma})$ interpolated with
- context-sensitive discriminative (MaxEnt) model $P(\text{lemma} | \text{src\_lemma}, \text{other features})$

He agreed with the unions to cut all overtime. Dohodl se s odbory na zrušení všech přesčasů.
2014: VowpalWabbit-based transfer

- VW is an ultra-fast and modular machine learning toolkit
- optimized SGD (AdaGrad, dense+sparse features, ...)
- cost-sensitive one-against-all reduction to binary classification
- logistic loss enables probabilistic interpretation (for HMTM)
- all lemmas in one model, fixed memory requirements
- label-dependent features (features shared for more lemmas)
VowpalWabbit Example

training data:

shared |S lemma=start formeme=v;for+ger neg=neg1 tag=VBG ...
1:0 _začít#V |T start^začít#V |P start^V
2:1 _zahájení#N |T start^zahájení#N |P start^N
3:1 _začínat#V |T start^začínat#V |P start^V
...
21:1 _spouštění#N |T start^spouštění#N |P start^N

training command:

vw -d train.data -c -f my.model
--loss_function=logistic --csoaa_ldf=mc -b 29 -qST
--holdout_off --passes 1 -l 3

test command:

vw -d test.data -c -i my.model -t -r out.predictions
Future plans

- non-isomorphic transfer
- experiments with VowpalWabbit
- include word embeddings (word2vec) as features
  - of the translated word (for rare words)
  - of its dependency context (for ambiguous words)
  - plus target-language embeddings of the translation
  - NN with a hidden layer
Thank you