Treebank Translation

Daniel Zeman, Rudolf Rosa

April 18, 2024
Closely Related Languages: Lexicalized Direct Transfer

  - In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pp. 210–219, Valencia, Spain

- Data from UD 1.4
  - Czech → Slovak
  - Slovenian → Croatian
  - Danish, Swedish → Norwegian
VarDial 2017 Baseline

- UDPipe, no parameter optimization
- Target tags predicted by UDPipe (supervised model!)

<table>
<thead>
<tr>
<th>Target</th>
<th>Source</th>
<th>DlxUAS</th>
<th>DlxLAS</th>
<th>LexUAS</th>
<th>LexLAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slovak</td>
<td>Czech</td>
<td>60.68</td>
<td>48.91</td>
<td>65.70</td>
<td>53.72</td>
</tr>
<tr>
<td>Croatian</td>
<td>Slovenian</td>
<td>62.64</td>
<td>50.81</td>
<td>63.94</td>
<td>53.35</td>
</tr>
<tr>
<td>Norwegian</td>
<td>Danish</td>
<td>65.23</td>
<td><strong>55.17</strong></td>
<td>64.53</td>
<td>54.91</td>
</tr>
<tr>
<td>Norwegian</td>
<td>Swedish</td>
<td>66.96</td>
<td><strong>57.54</strong></td>
<td>66.24</td>
<td>56.63</td>
</tr>
<tr>
<td>Norwegian</td>
<td>Danish+Swedish</td>
<td>68.58</td>
<td>58.80</td>
<td>69.02</td>
<td><strong>59.95</strong></td>
</tr>
</tbody>
</table>
Recall from Delex: Danish – Swedish Setup

  - In *IJCNLP 2008 Workshop on NLP for Less Privileged Languages*, pp. 35–42, Hyderabad, India

- CoNLL 2006 treebanks (dependencies)
  - Danish Dependency Treebank
  - Swedish Talbanken05

- Two constituency parsers:
  - “Charniak”
  - “Brown” (Charniak N-best parser + Johnson reranker)

- Other resources
  - JRC-Acquis parallel corpus
  - Hajič tagger for Swedish (PAROLE tagset)
Recall from Delex: Danish – Swedish Setup

- Other resources
  - JRC-Acquis parallel corpus
  - Did not need it for delex. But...
Acquis is a parallel corpus
  More than 430,000 sentences
GIZA++ & lexical weighting generate da-sv glossary
Glosses

- Acquis is a parallel corpus
  - More than 430,000 sentences
- GIZA++ & lexical weighting generate da-sv glossary
- Always use highest weighted gloss
- Translate Swedish word-by-word to Danish
Acquis is a parallel corpus
  - More than 430,000 sentences
GIZA++ & lexical weighting generate da-sv glossary
Always use highest weighted gloss
Translate Swedish word-by-word to Danish
Use Danish parser
Many unknown words are known now!

Translated target to source
### Most Frequent da / sv Words

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>0.024</td>
<td>och</td>
<td>0.027</td>
</tr>
<tr>
<td>og</td>
<td>0.024</td>
<td>att</td>
<td>0.027</td>
</tr>
<tr>
<td>at</td>
<td>0.021</td>
<td>i</td>
<td>0.021</td>
</tr>
<tr>
<td>er</td>
<td>0.017</td>
<td>är</td>
<td>0.018</td>
</tr>
<tr>
<td>en</td>
<td>0.014</td>
<td>som</td>
<td>0.017</td>
</tr>
<tr>
<td>til</td>
<td>0.013</td>
<td>en</td>
<td>0.015</td>
</tr>
<tr>
<td>af</td>
<td>0.013</td>
<td>det</td>
<td>0.013</td>
</tr>
<tr>
<td>det</td>
<td>0.012</td>
<td>av</td>
<td>0.012</td>
</tr>
<tr>
<td>på</td>
<td>0.012</td>
<td>på</td>
<td>0.011</td>
</tr>
</tbody>
</table>
Denne forordning træder i kraft den 1. marts 1986 med forbehold af ikrafttrædelse af traktaten vedrørende Spaniens og Portugals tiltrædelse.

Denna förordning träder i kraft den 1 mars 1986 under förutsättning att Anslutningsakten för Spanien och Portugal träder i kraft.
Bestemmelserne i denne aftale kan ændres og revideres helt eller delvis efter fælles overenskomst mellem parterne.

Bestämmelserna i detta avtal får ändras eller revideras helt eller delvis efter gemensam överenskommelse mellan parterna.
1. A Contracting Party may **terminate** this Convention by written notification to the Depositary.
behandlingsaktörer  behandlingsvirksomheder
behandlingsanläggning  behandlingsanlæg
behandlingsanläggningar  behandlingsvirksomheders
behandlingsanläggningen  behandlingsanlægget
behandlingsdatum  datøn
behandlingsformer  behandlingsmuligheder
behandlingsfrister  frister
behandlingsförfaranden  behandlingsprocedurer
behandlingsförsök  befolkningsforsøg
behandlingsindikation  indikation
behäftad  behæftet
behåll  behold
Unlabeled F Scores

- da-da lexicalized: Charniak = 78.16, Brown = 78.24
  - (CoNLL train 94K words, test 5852 words)
- sv-sv lexicalized: Charniak = 77.81, Brown = 78.74
  - (CoNLL train 191K words, test 5656 words)
- da-sv lexicalized: Charniak = 43.28, Brown = 41.84
  - (no morphology tweaking)
- da-da delexicalized: Charniak = 79.62, Brown = 80.20
  - (hybrid sv-da Hajič-like tagset = “words”, Penn POS = “tags”)
- sv-sv delexicalized: Charniak = 76.07, Brown = 77.01
- da-sv delexicalized: Charniak = 65.50, Brown = 66.40
Unlabeled F Scores

- da-da lexicalized: Charniak = 78.16, Brown = 78.24
  - (CoNLL train 94K words, test 5852 words)
- sv-sv lexicalized: Charniak = 77.81, Brown = 78.74
  - (CoNLL train 191K words, test 5656 words)
- da-sv lexicalized: Charniak = 43.28, Brown = 41.84
  - (no morphology tweaking)
- da-da delexicalized: Charniak = 79.62, Brown = 80.20
  - (hybrid sv-da Hajič-like tagset = “words”, Penn POS = “tags”)
- sv-sv delexicalized: Charniak = 76.07, Brown = 77.01
- da-sv delexicalized: Charniak = 65.50, Brown = 66.40
- da-sv glossed: Charniak = 63.40, Brown = 61.50
Glosses with Self-Training

- Danish treebank
  - PARSER 0
  - RERANKER
    - DELEX
    - GLOSSES
      - Swedish Acquis
        - PARSER 1
          - RESTUFF
            - Parsed Swedish Acquis
  - Swedish test
Unlabeled F Scores

- da-da lexicalized: Charniak = 78.16, Brown = 78.24
  - (CoNLL train 94K words, test 5852 words)
- sv-sv lexicalized: Charniak = 77.81, Brown = 78.74
  - (CoNLL train 191K words, test 5656 words)
- da-sv lexicalized: Charniak = 43.28, Brown = 41.84
  - (no morphology tweaking)
- da-da delexicalized: Charniak = 79.62, Brown = 80.20
  - (hybrid sv-da Hajič-like tagset = “words”, Penn POS = “tags”)
- sv-sv delexicalized: Charniak = 76.07, Brown = 77.01
- da-sv delexicalized: Charniak = 65.50, Brown = 66.40
- da-sv glossed: Charniak = 63.40, Brown = 61.50
- da-sv glossed+self: Charniak = 64.48, Brown = 63.32


Also addresses unclear points of Hwa et al. (2004)
Treebank Translation

- Translate source treebank to target language
- Extract word alignments
  - Either directly from Moses / MT system (better!)
  - Or afterwards, using GIZA++ / Fastalign etc. (worse)
    - Extra noise from separate alignment
    - Treebank too small to compute alignment
    - Added parallel data ⇒ noise, domain?

- Comparison with treebank projection across alignments:
  - Avoid double noise (parse source side + project)
  - Avoid domain shift (source treebank vs. parallel corpus)
    - There is still domain shift between source and target treebank, if it exists.
  - Machine translation more literal than human ⇒ better alignment
Zeman and Resnik (2008) glosses (taken from GIZA++)

Tiedemann (2014): force Moses to use 1-word phrases
  • He uses it to project POS-tag models (no DUMMY words on target side)

Rosa et al. (2017), VarDial
  • Closely related languages
  • Use this even for trees!
  • Projection of relations is straightforward
  • Unknown words:
    • Leave as is
    • Or learn a character-based “transcription” model?
Summary of Cross-lingual Parsing

- **Direct transfer**
  - Train on source language, apply to target language
    - Delexicalized
    - Lexicalized

- **Treebank projection across parallel data**
  - Train on source language, parse source side of parallel data
  - Project trees to target side of parallel data
  - Train a target language parser on that

- **Treebank translation**
  - Train an MT system on parallel data
  - Enforce word-by-word MT, or extract alignments afterwards
    - Either: Translation of training trees from source to target
    - Or: Translation of test sentences from target to source

https://ufal.cz/courses/npfl1120
**Levenshtein Distance**


- Minimum number of character edits to get from string $a$ to $b$

- Edit operations:
  - Insert a character
  - Delete a character
  - Substitute a character for another character

- ⇒ learn context-sensitive edits between languages $A$ and $B$
  - E.g. Czech to Slovak:
    - $pro$ → $pre$; $při$ → $pri$; $-ní$ → $-nie$...