Treebank Translation

Daniel Zeman, Rudolf Rosa

April 17, 2020



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

Closely Related Languages: Lexicalized Direct Transfer

- Rudolf Rosa, Daniel Zeman, David Mareček, Zdeněk Žabokrtský (2017). Slavic Forest, Norwegian Wood.
  - 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!)

Target	Source	DIxUAS	DIxLAS	LexUAS	LexLAS
Slovak	Czech	60.68	48.91	65.70	**53.72**
Croatian	Slovenian	62.64	50.81	63.94	**53.35**
Norwegian	Danish	65.23	**55.17**	64.53	54.91
Norwegian	Swedish	66.96	**57.54**	66.24	56.63
Norwegian	Danish+Swedish	68.58	58.80	69.02	**59.95**

Recall from Delex: Danish – Swedish Setup

- Daniel Zeman, Philip Resnik (2008). Cross-Language Parser Adaptation between Related Languages
  - 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...

Glosses

- 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

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
- Use Danish parser
- Many unknown words are known now!

- Translated target to source

Most Frequent da / sv Words

i	0.024	och	0.027
og	0.024	att	0.027
at	0.021	i	0.021
er	0.017	är	0.018
en	0.014	som	0.017
til	0.013	en	0.015
af	0.013	det	0.013
det	0.012	av	0.012
på	0.012	på	0.011

Aligned Sentences 1

- 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.

Aligned Sentences 2

- 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.

Aligned Sentences 3

- 1. Enhver kontraherende part kan opsige denne konvention ved skriftlig henvendelse til depositaren.
- 1. En fördragsslutande part fåt säga upp denna konvention genom skriftlig notifikation till depositarien.
- 1. A Contracting Party may terminate this Convention by written notification to the Depositary.

Excerpt from sv-da Glossary

behandlingsaktörer	behandlingsvirksomheder
behandlingsanläggning	behandlingsanlæg
behandlingsanläggningar	behandlingsvirksomheders
behandlingsanläggningen	behandlingsanlægget
behandlingsdatum	datøn
behandlingsformer	behandlingsmuligheder
behandlingsfrister	frister
behandlingsförfaranden	behandlingsprocedurer
behandlingsförsök	befolkningsforsøg
behandlingsindikation	indikation
behäftad	behæftet
behåll	behold

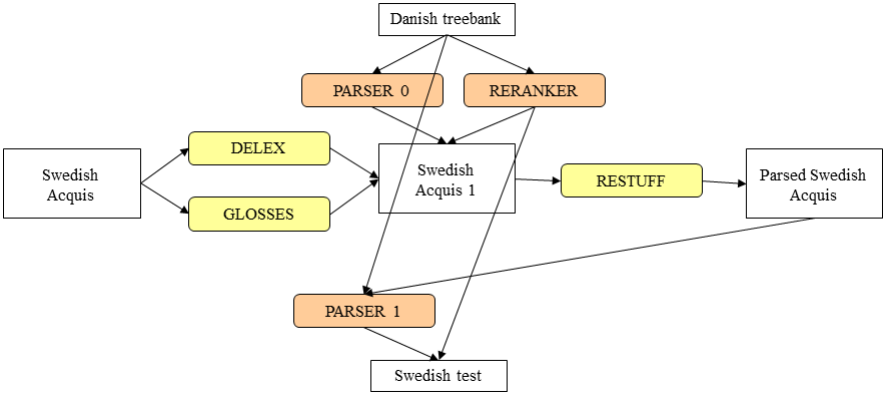
Unlabeled F Scores

- 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



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

Treebank Translation

- Jörg Tiedemann (2014). Rediscovering Annotation Projection for Cross-Lingual Parser Induction.
  - In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1854–1864, Dublin, Ireland
- 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 $\Rightarrow$ noise, domain?

- 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 $\Rightarrow$ better alignment

Word-by-Word Translation

- 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!
  - $\Rightarrow$ Projection of relations is straightforward
  - Unknown words:
    - Leave as is
    - Or learn a character-based "transcription" model?

Levenshtein Distance

- Vladimir Levenštejn (1965). Двоичные коды с исправлением выпадений, вставок и замещений символов [Binary codes capable of correcting deletions, insertions, and reversals]. Доклады Академий Наук СССР. 163 (4): 845–8.

- 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

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