The UdS POS Tagging Systems @ EmpiriST 2015

Jakob Prange Andrea Horbach Stefan Thater

Saarland University, Saarbrücken

3rd NLP4CMC Workshop, KONVENS 2016, Bochum

UNIVERSITÄT DES SAARLANDES

Projekt Schreibgebrauch
Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen

Table of Contents

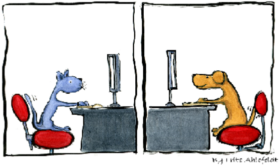
Revisiting our Submissions

- Certain phenomena occur over and over in CMC
- Adding in-domain training data will help to cope with them

1st approach: "Retrain" [Horbach et al. 2014, 2015]

TIGER + EmpiriST training set + "Schreibgebrauch" project training set

+



Revisiting our Submissions

- Certain phenomena occur over and over in CMC
- Adding in-domain training data will help to cope with them

1st approach: "Retrain" [Horbach et al. 2014, 2015]

TIGER + EmpiriST training set + "Schreibgebrauch" project training set

+

Revisiting our Submissions

- Certain phenomena occur over and over in CMC
- Adding in-domain training data will help to cope with them

1st approach: “Retrain” [Horbach et al. 2014, 2015]

Combine TIGER + EmpiriST training set + “Schreibgebrauch” project training set (boosted 5 times)

+

Revisiting our Submissions

1st approach: "Retrain" [Horbach et al. 2014, 2015]

TIGER + EmpiriST training set + "Schreibgebrauch" project training set (boosted 5 times)

Pros

✓ big performance boost

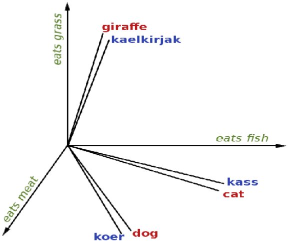
Cons

× many words still not in training data

× expensive to annotate more data

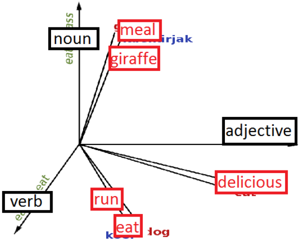
Revisiting our Submissions

- unsupervised learning
- profit from large, raw in-domain data set
- assumption: words have the same POS tags as their distributional neighbours



Revisiting our Submissions

- unsupervised learning
- profit from large, raw in-domain data set
- assumption: words have the same POS tags as their distributional neighbours



Revisiting our Submissions

- unsupervised learning
- profit from large, raw in-domain data set
- assumption: words have the same POS tags as their distributional neighbours

2nd approach: "Distributional" [Prange et al. 2015]

for each unkown word type:

- generate known candidates based on distributional similarity
- rank POS tags of candidates
- propose highest ranked POS tag(s) to the tagger

Revisiting our Submissions

2nd approach: "Distributional" [Prange et al. 2015]

for each unkown word type:

- generate known candidates based on distributional similarity
- rank POS tags of candidates
- select highest ranked POS tag(s) to the tagger

Pros

✓ no additional manual annotation

✓ covers more words

Cons

× local context not considered

→ multiple readings of one word cannot be distinguished (only indirectly via off-the-shelf tagging software)

Revisiting our Submissions

- assumption: unknown words are often misspellings and similar to their intended forms

3rd approach: “Surface”

for each unkown word type:

- generate candidates based on string-similarity

for each unkown word token:

- rank candidates in context by language model
- replace unknown word with highest ranked candidate

Revisiting our Submissions

- assumption: unknown words are often misspellings and similar to their intended forms

Revisiting our Submissions

3rd approach: "Surface"

for each unkown word type:

- generate candidates based on string-similarity

for each unkown word token:

- rank candidates in context by language model
- replace unknown word with highest ranked candidate

Pros

✓ no additional manual annotation

✓ local context considered

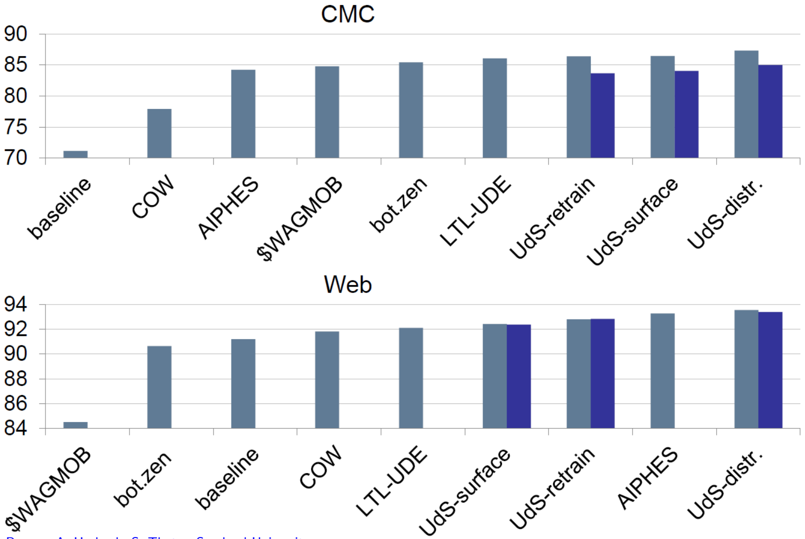
Cons

× very small performance boost, if any

× "overcorrection": not only typos, but also lexical gaps are replaced

Revisiting our Submissions

[% accuracy]



Afterthoughts and Ideas for the Future

- influence of data vs influence of algorithm

Afterthoughts and Ideas for the Future

- influence of data vs influence of algorithm
- oracle experiment shows there is room for improvement with an ideally combined system

Afterthoughts and Ideas for the Future

- influence of data vs influence of algorithm
- oracle experiment shows there is room for improvement with an ideally combined system
- new particle tags are problematic – also for humans? Would it help to re-annotate TIGER with STTS 2.0?

Afterthoughts and Ideas for the Future

- influence of data vs influence of algorithm
- oracle experiment shows there is room for improvement with an ideally combined system
- new particle tags are problematic – also for humans? Would it help to re-annotate TIGER with STTS 2.0?
- action words are problematic – (morphological) preprocessing? Tokenisation?

Thank you!

software available under
http://www.coli.uni-saarland.de/projects/schreibgebrauch/de/page.php?id=resources