FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

NLP for German CMC Data

Thomas Proisl · Philipp Heinrich

Lehrstuhl für Korpus- und Computerlinguistik, FAU Erlangen-Nürnberg

Data Set

EmpiriST 2015 gold standard

- Shared task on automatic linguistic annotation of computer-mediated communication (CMC) and web corpora (Beißwenger et al. 2016):
  - **CMC**: tweets, social and professional chats, comments, wiki talk pages
  - **Web**: web sites, blogs, Wikipedia articles, Wikinews

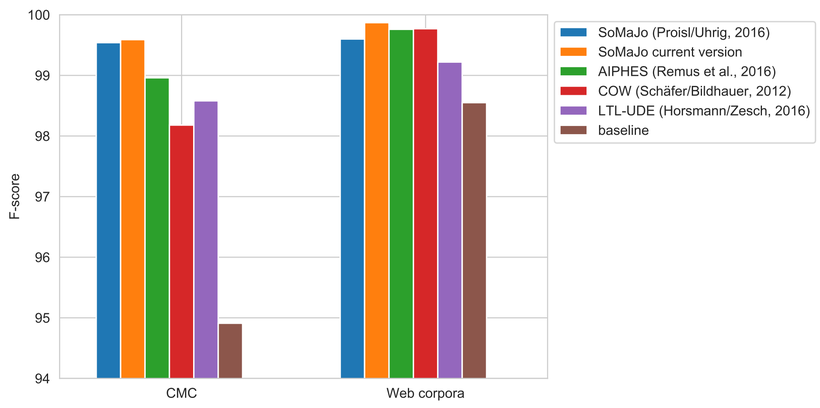
	CMC	Web
Training	5,109	4,944
Test	5,237	7,568
Total	10,346	12,512

- Manually tokenized and annotated with STTS_IBK
  - STTS + 18 additional tags (Beißwenger et al. 2015)
- Manually normalized and lemmatized (Proisl et al. forthcoming)

Tokenization

Successful rule-based approaches

- Even a simple baseline (whitespace tokenizer that splits off punctuation) works surprisingly well
- Best-performing tokenizers achieve F_1 scores > 0.99
- No need for ML techniques



Results for tokenization (F_1 scores)

Lemmatization

New gold standard

- Two lemmatization strategies:
  - Surface-oriented lemmatization (based on inflectional suffixes, retains non-standard orthographical features)

    Grigfe → *Grigf*
  - Normalized lemmatization (correct obvious spelling errors, standard form of non-standard tokens)

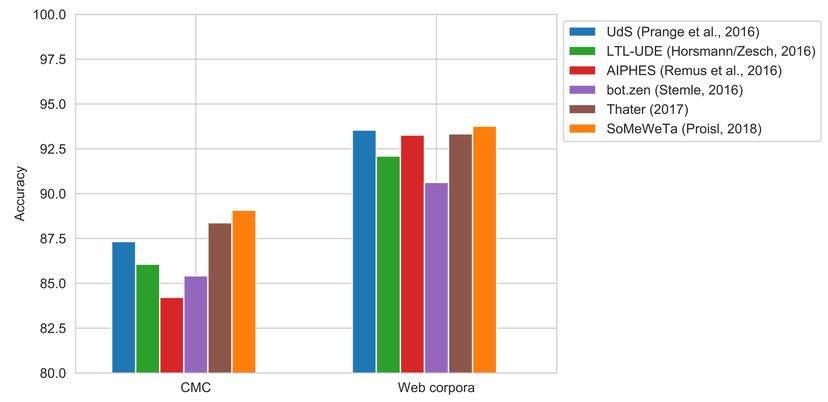
    Grigfe → *Griff*
- Four student annotators, unclear cases decided in group meetings with supervisors
- Inter-annotator agreement (Cohen's κ): 0.93–0.97
- Baselines (accuracy, ignoring case):
  - Do-nothing: Always return the word form
  - Weak: Given word form and POS, return most frequent lemma
  - Strong: Apache OpenNLP maximum entropy lemmatizer

Baseline	surface-oriented	normalized
Do-nothing	71.63	70.73
Weak	83.90	83.19
Strong	87.50	85.97
Human avg.	94.70	94.40

Part-of-Speech Tagging

Various ML techniques

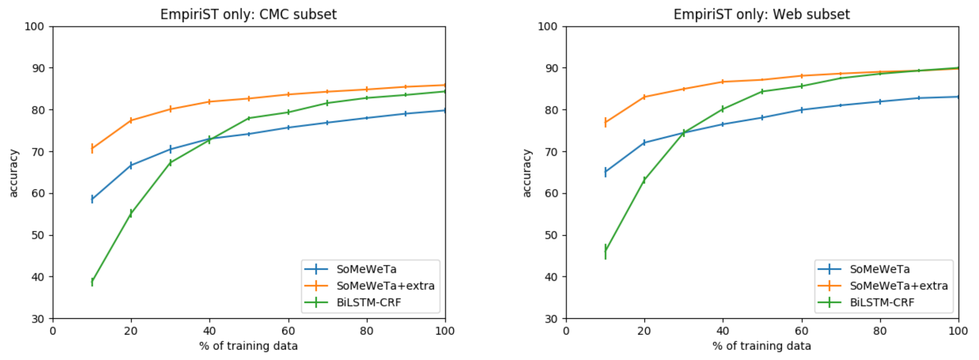
- HMM (UdS, Thater 2017)
- CRF (AIPHES, LTL-UDE)
- LSTM (bot.zen)
- averaged structured perceptron (SoMeWeTa)

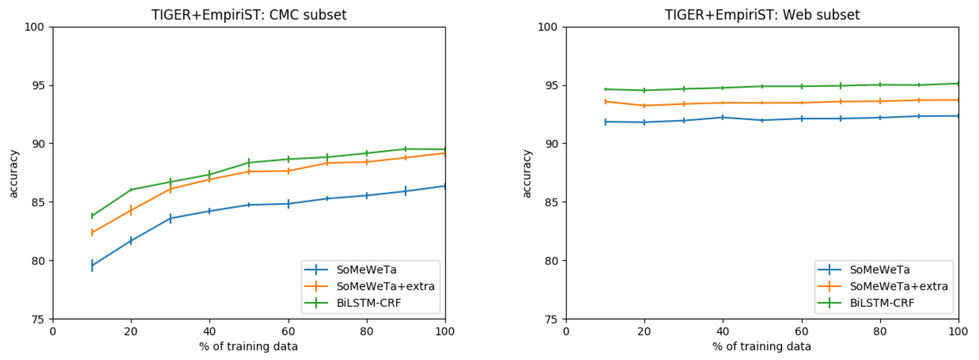


Results for part-of-speech tagging (accuracy)

Further Experiments

- Aim: Compare best-performing system (SoMeWeTa) to state-of-the-art BiLSTM-CRF tagger that uses word- and character-level BiLSTMs (Riedl and Padó 2018)
- Setting:
  - Only EmpiriST training data vs. additional pretraining on TIGER
  - SoMeWeTa with and without external resources
  - BiLSTM-CRF tagger with pretrained word embeddings
- External Resources and Transfer Learning
  - SoMeWeTa: Coarse-grained word class information from Morphy (Lezius 2000), Brown clusters from DECOW14
  - BiLSTM-CRF: Pretrained fastText embeddings





Learning curves

- SoMeWeTa: Additional resources lead to improvements (6–12 points); graceful degradation
- BiLSTM-CRF: Steeper learning curve; probably outperforms SoMeWeTa (even with additional resources) with slightly more training data
- Transfer learning leads to better results (4–9 points)
- Web corpus very similar to TIGER → very flat learning curve
- Learning curve for CMC not flattened out
- BiLSTM-CRF outperforms SoMeWeTa (0.3–1.4 points), parallel learning curves

http://www.linguistik.fau.de https://github.com/fau-klue/empirist-corpus {thomas.proisl,philipp.heinrich}@fau.de