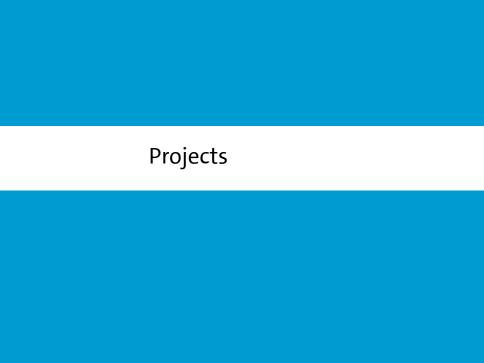


Benjamin Milde, Prof. Dr. Chris Biemann

DEEP LEARNING FOR LANGUAGE AND SPEECH - SEMINAR PROJECTS







#1 GermaNER tagger (1)

- German named entity recognition (organizations, names, places) with a twist - can you make the model compact?
- Dataset: https://github.com/tudarmstadt-lt/GermaNER
- Model accuracy vs. model size
- Many OOVs, will need a character model
- Might be used by LT projects e.g. new/s/leak (NetWork of Searchable Leaks), https://www.inf.uni-hamburg.de/en/ inst/ab/lt/resources/demos/new-s-leak.html



Projects 0 • 0 0 0 0 0 0 0 0

#1 GermaNER tagger (2)

```
Schartau B-PER
sagte 0
dem O
" 0
Tagesspiegel B-ORG
" 0
vom O
Freitag O
.0
Fischer B-PER
sei O
" N
... n
Firmengründer O
Wolf B-PER
Peter I-PER
Bree I-PER
arbeitete O
Anfang O
der O
```

#2 Multilingual Emoji Prediction

- SemEval-2018 task 2 competition
- https://competitions.codalab.org/competitions/ 17344#learn_the_details



#3 Irony detection in English tweets

- SemEval-2018 task 3 competition
- competitions.codalab.org/competitions/17468
- Task a) binary classification:
 - I just love when you test my patience!! #not Had no sleep and have got school now #not happy
- Task b)
- i) verbal irony realized through a polarity contrast, ii) verbal irony without such a polarity contrast (i.e., other verbal irony), iii) descriptions of situational irony, iv) non-irony.

#4 One Million Posts Corpus

- DER STANDARD is an Austrian daily broadsheet newspaper.
- The data set contains a selection of user posts from the 12 month time span from 2015-06-01 to 2016-05-31.
- There are 11,773 labeled and 1,000,000 unlabeled posts in the data set.
- Labels: Sentiment (negative/neutral/positive), Off-Topic (yes/no), Inappropriate (yes/no), Discriminating (yes/no)



#5 Common voice

- Recently released speech dataset from mozilla
- 20000 speakers, some information about age, gender and accent available
- Predict those attributes (separately or together)

```
['cv-valid-dev/sample-004039.mp3', 'are you enjoying london',
   '1', '0', 'thirties', 'female', 'england', '']
['cv-valid-dev/sample-004040.mp3', 'they placed the symbols of the pilgrimage on the doors of their house
'3', '0', 'sixties', 'female', 'us', '']
['cv-valid-dev/sample-004041.mp3', 'he could always go back to being a shepherd',
'3', '0', '', '', '', '']
['cv-valid-dev/sample-004042.mp3', 'its lower end was still embedded',
'6', '0', 'twenties', 'female', '', '']
['cv-valid-dev/sample-004043.mp3', "i'm going into the desert the man answered turning back to his reading
'1', '0', '', '', '', '']
['cv-valid-dev/sample-004044.mp3', 'as the sun rose the men began to beat the boy',
'2', '0', 'thirties', 'male', 'australia', '']
['cv-valid-dev/sample-004045.mp3', 'i have to find a man who knows that universal language',
'3', '0', 'thirties', 'male', 'us', '']
['cv-valid-dev/sample-004046.mp3', 'the picnic was ruined by a marching band',
'1', '0', 'twenties', 'male', 'scotland', '']
```

#6 Computational Paralinguistics

- http://emotion-research.net/sigs/speech-sig/ is2017_compare.pdf
- Normal speech vs. speaker has a cold challenge
- ALC corpus: https: //clarin.phonetik.uni-muenchen.de/BASRepository/
- Intoxicated vs normal speech.
- Try transfer or multitask learning!

- abbreviate → AH B R IY V IY EY T
- Sequence to sequence models
- Idea: Combine NLP and speech
- Cmudict and speech samples (dict.cc)
- http://www.speech.cs.cmu.edu/cgi-bin/cmudict
- dict.cc



#8 Your own idea here

- As long as it is about text/speech
- And doable in a short time