Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Benjamin Milde, Prof. Dr. Chris Biemann

DEEP LEARNING FOR LANGUAGE AND SPEECH - SEMINAR PROJECTS

Projects

Focus this year: Recurrent architectures

- Three main task categories:
- Classification
- Sequence labeling
- Sequence-to-sequence
- This is also the order of technical/programming difficulty.

#1 OffensEval 2019

- Felix, Christina, Vadym
- Identify offense, aggression, and hate speech in user-generated content
- Sub-task A - Offensive language identification;
- Sub-task B - Automatic categorization of offense types;
- Sub-task C - Offense target identification.
- SemEval challenge `https://competitions.codalab.org/competitions/20011`

#1 OffensEval 2019 - examples

```
90194 @USER @USER Go home you're drunk!!! @USER #MAGA #Trump2020  URL
77294 @USER @USER She is an ugly black hearted troll URL
72024 @USER Why? Why are liberals so trashy?
32322 @USER @USER @USER @USER And you're just another Twitter asshole.
14116 @USER who's the loser bitch!  lol you! fuck you #MAGA
34301 @USER @USER FASCISTS OF THE FUTURE WILL CALL THEMSELVES ANTI
FASCIST.  ANTIFA COMES TO MIND.
```

#2 Insincere Questions (Quora)

- Jaakko, David
- `https://www.kaggle.com/c/quora-insincere-questions-classification`
- An insincere question is defined as a question intended to make a statement rather than look for helpful answers.
- Kaggle challenge: 1st Place - $12,000, 2nd Place - $8,000, 3rd Place - $5,000
- 2 months to go
- Binary classification task

#2 Insincere Questions Examples

If both Honey Singh and Justin Bieber fall from the 5th floor, who will survive?

Why are there so many sensitive liberals on Quora?

What's so great about Singapore?
It's expensive, it rains every day, there's conscription and you have to work 9-10 hours a day.

Why do all model thin girls desperately want kids?

#3 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
- Sequence labeling task

#3 GermaNER tagger (2) - dataset

```
Schartau B-PER
sagte O
dem O
" O
Tagesspiegel B-ORG
" O
vom O
Freitag O
,O
Fischer B-PER
sei O
" O
... O

Firmengründer O
Wolf B-PER
Peter I-PER
Bree I-PER
arbeitete O
Anfang O
der O
```

#4 Emoji Prediction

- Max, Luca
- Preprocessed (2016): `http://ltdata1.informatik.uni-hamburg.de/smileytastic`
- For newer years, raw comment data in: `https://files.pushshift.io/reddit/comments/`
- Classification or Sequence-to-Sequence
- Remove smileys from sentences and predict them:

```
Did you by any chance bought the kit at Radio Shack?
;P


That book is really confusing.
There's a headline and then a link to somewhere else.
o_O

I can't tell if anything in this topic is sarcastic or not
```

#5 Text Normalization (I)

- Helge
- `https://www.kaggle.com/c/text-normalization-challenge-english-language`
- Text-to-speech synthesis (TTS) and automatic speech recognition (ASR), require text to be converted from written expressions into appropriate ”spoken” forms.
- This is a process known as text normalization, and helps convert 12:47 to ”twelve forty-seven” and $3.16 into ”three dollars, sixteen cents.”
- Sequence labeling task + Sequence-to-sequence

#5 Text Normalization (II) - dataset

```
A    <self>
baby    <self>
giraffe    <self>
is    <self>
6ft    six feet
tall    <self>
and    <self>
weighs    <self>
150lb    one hundred fifty pounds
.    sil
```

#5 Text Normalization (III) - dataset

```
Проверено       <self>
12 февраля 2013    двенадцатого февраля две тысячи тринадцатого года
,    sil
Архивировано    <self>
из    <self>
первоисточника    <self>
15 февраля 2013    февраля две тысячи тринадцатого года
.    sil
```

#6 G2P

- Hans, Inga, Tim
- abbreviate → AH B R IY V IY EY T
- Sequence to sequence models
- The CMU Pronouncing Dictionary
- `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`
- German dictionary: `https://raw.githubusercontent.com/marytts/marytts-lexicon-de/master/modules/de/lexicon/de.txt`
- Sequence to sequence task

#7 Writing prompts

- Patrick, Dennis
- Generative model, Sequence to sequence task, e.g.:

Input: The year is 1910. Adolf Hitler, a struggling artist, has fought off dozens of assasination attemps by well meaning time travelers, but this one is different. This traveller doesn't want to kill Hitler, he wants to teach him to paint. He pulls off his hood to reveal the frizzy afro of Bob Ross.

#7 Writing prompts

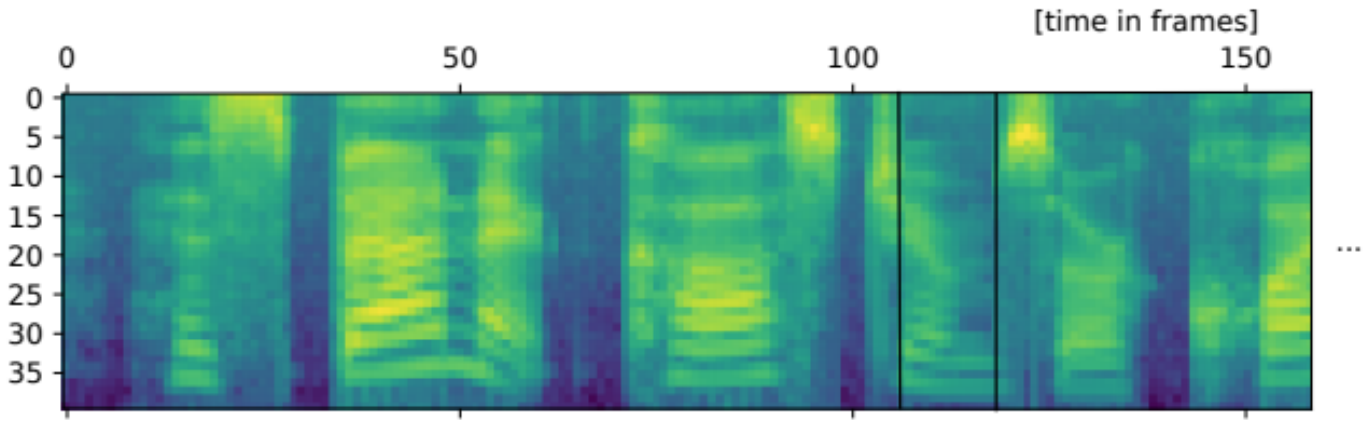
There he sat, twirling his personal, stylized mustache. It was avant garde, just like he wanted to be. The man, as he was so, just wanted a place in this world for his art. He continues to stare at the easel, thinking.
After a while he felt a firm, calming hand on his shoulder. He sighed, hanging his head wearily. ”Are you yet another man come to end my life, if you can even see it that way?” The hand didn’t answer, as it had no mouth. However, it’s owner did, speaking the soft, assuaging tones that had come to make him famous.
”No sir. I’ve seen too much death and war to want to do another such thing. Instead, I have come as a tutor. Here, grab that 2 inch brush and dip it in some titanium white and prussian blue.”
Hitler did such a thing, and the man behind him nodded. ”Good. Now, mix them together, until you have a rather nice pale blue...”

...

#8 Speech tokenizer

- Florian, Julian
- Segment spoken language
- Word, phoneme or syllable boundaries
- Sequence labelling task

#8 Speech tokenizer



#9 Your own idea here

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

Timeline

- Find a project and a team (2 persons)
- Start small, reduce training data set size first!
- First use CPU, then if everything works, use GPUs
- Next week: Learn to use the Nvidia Tesla GPUs on the Hummel cluster (NVIDIA K80)
- We will have about one node (2 GPUs) per team
- Generate a public key: https://www.rrz.uni-hamburg.de/services/hpc/grundwissen/ssh-keys.html

Pretrained embeddings

- In NLP tasks with little training data it may make sense to use precomputed embeddings:
- GoogleNews-vectors-negative300 - `https://code.google.com/archive/p/word2vec/`
- glove.840B.300d - `https://nlp.stanford.edu/projects/glove/`
- paragram_300_sl999 - `https://cogcomp.org/page/resource_view/106`
- wiki-news-300d-1M - `https://fasttext.cc/docs/en/english-vectors.html`

Extra: 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 houses
 '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', '']
```

Extra: 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)