Natural Language Processing (PA153)

Pavel Rychlý

Summary

Problems with NLP

- Problems with NLP
  - Zipf's law
  - Ambiguity
  - Variability
- Approaches
  - symbolic (rule-based)
    - no data available
  - statistical
  - neural (deep learning)
    - huge data available

Statistical NLP

- counts
- keywords
- collocations, multi-word units
- language modeling

Language Modeling

- probability of senteces, chain rule
- n-grams, Markov's assumption
  $p(W) = \prod_i p(w_i|w_{i-2}, w_{i-1})$
- maximum-likelihood estimation gives zero probabilities
- smoothing
- evaluation using cross entropy, perplexity

Text Classification

- applications
- Naive Bayes Classifier
- evaluation:
  - precision
  - recall
  - accuracy

Continuous Space Reprasentation

- words as vectors, word embeddings
- methods of learning vectors
- evaluation of words embeddings
- *optional homework: Stability of word embeddings*

Neural Networks

- structure of NN
- matrix representation
- activation functions
- NN training
  - stochastic gradient descent
  - backpropagation
- sub-word tokenization
  - *opt. hw: subword coverage*

Recurrent NN

- language modeling using NN
- training RNN
- problems in training RNN
- LSTM
- Bidirectional, multi layer RNN

Simple NLP using NN

- Named Entity Recognition (NER)
- language modeling
- training
- evaluation
- *opt. hw: NN for adding accents*

Machine translation

- sequence to sequence RNN
- attention
- decoding, beam search
- MT evaluation: BLEU

Transformers

- encoder, decoder
- encoding positon
- attention structure

Pretrained models

- Encoder only
- Decoder only
- Encoder-decoder
- training strategies
- BERT, GPT, T5

Question Answering

- QA types
- usage
- reading comprehension
- applying NN for QA

Lexicography

- current trends in lexicography
- lexical database
- data processing
- dictionary writing systems

Recipe for Training NN

- NN training fails silently

1. Become one with the data
2. Set up the end-to-end training/evaluation skeleton + get dumb baselines
3. Overfit
4. Regularize
5. Tune
6. Squeeze out the juice