MUNI
FI

Natural Language Processing

Summary

Pavel Rychlý

16 Dec 2024

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

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, ChrF++

Transformers

- encoder, decoder
- encoding positon
- attention structure

Pretrained models

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

Question Answering

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

Where to start

- Hugging Face
  - models
    - code
    - pre-trained, ready to use
  - datasets
    - sometimes with evaluaton
- transformers library
  - very complex
  - 3 implementations: Jax, PyTorch, TensorFlow

Pre-trained models

- olamma
- llama.cpp
  - run the LLaMA model using 4-bit integer quantization on a MacBook
- optimizations
  - float16, bfloat16
  - quantization

Training from scratch

- nanoGPT
  - easy to read
  - minimal dependencies
- nanoT5
  - train T5 on 1xA100 GPU in less than 24 hours
- BabyLM Challenge
  - 100/10M word text-only dataset