Foundations of Natural Language Processing

Revision and Q&A

Mirella Lapata and Ivan Titov

March 25, 2025

School of informatics

Recall: Class Goals

- Our focus was on the core concepts and methods behind language technologies — the building blocks for future applications.
  - Key linguistic phenomena and challenges
  - Classic (pre-deep learning) approaches
  - Modern deep learning methods, including an intro to LLMs

- . . . while more advanced topics are explored in later courses:
  - Natural language understanding, generation, and machine translation (NLU+)
  - Automatic speech recognition

Today

- We will go briefly over the material and highlight some topics and ideas . . .
  - Important: do not assume that if we have not highlighted a certain topic, it is not going to be examined
- Ask questions about the class

Revision – Exam

- Previous year exams can be helpful – but keep in mind:
  - This year's exam is closed-book (no materials allowed)
  - Expect some **bookwork-style questions**
  - We've modernized the course — the 2024/25 content differs significantly from 2023/24, and even more from 2022/23
  - For common topics, exam questions from previous years are still relevant/useful.

Revision – Study Resources

- Do readings from Jurafsky & Martin (and other sources we pointed out)
  - We won't examine what has not been at all discussed in the class . . .
  - but the book covers much of it in more detail, with more examples. . .
    - It will definitely help you prepare better for the exam
- Note though that there is material which is not covered in J&M
- Use lecture slides / recordings
- Use quizzes and tutorials

Be able to Deal with Novel Problems

- Methods are your toolbox
- Be prepared to reduce a given problem to the **modeling** setups we discussed
  - How can we convert a problem into a (set of) classification problems? Sequence labelling problem?
- The same applies to algorithms, evaluation, . . .
- Expect to be able to write down some formulas from memory and to draw different components of model architectures.
- In all cases think of input, output, training objective, why does your solution make sense for the task?

Ambiguities and challenges in NLP

- **Remember different challenges for NLP** (e.g., sparsity, ambiguity, robustness, . . .)
  - Be ready to identify the key challenges for a given specific setting
  - and ways to circumvent these challenges

- **Ambiguity**
  - A topic we discussed a lot in the class
  - Why is it a problem? How to deal with it? Types of ambiguities?

Corpora and Experimental Design

- Make sure you understand challenges and considerations in designing a corpus collection and how these have changed for LLMs
  - E.g., be ready to discuss the differences between collections with annotation and those without.

- Make sure you understand how to evaluate different types of NLP models and hypotheses
  - E.g., you can be asked to consider a specific setting/application and *come up with* a way of evaluating NLP tools, or discuss advantages/disadvantages of alternatives
  - We talked about evaluation in the context of classification (accuracy, F1), parsing (bracket F1), LLMs (perplexity), and seq2seq models (n-gram overlap metrics, BLEU)

Text Classification Methods

- **Naïve Bayes and Logistic Regression**
  - Parameter estimation and inference (i.e., how they're used)
  - Strengths and limitations
- **Neural Approaches**
  - Relation to logistic regression
  - Bag-of-embeddings models
  - Recurrent Neural Networks (RNNs)
    - Vanilla RNNs (covered in detail)
    - LSTMs and GRUs (high-level understanding)
  - Differences in expressivity; multi-layer and bidirectional architectures

Distributional Semantics / Word Embeddings

- Understand the underlying assumptions (what it can/can't do)
- Count-based methods and SVD
- Differences and similarities between sparse and dense embeddings
- Neural embeddings
  - **Skip-gram, including negative sampling**
  - Importance of negative sampling as an example of self-supervision

Language Models

- **N-gram language models**
  - How to estimate? How to evaluate? Limitations?
- Uses of language models to estimate probabilities and generate language
- We briefly talked about smoothing, you should understand the concept and why it is necessary for count-based LLMs
- **Neural language models**
  - Reduction to classification
  - Relation to smoothing in n-gram models
  - Estimation
- **Evaluation of language models**
- **Decoding**: greedy, sampling with temperature, top-k, nucleus sampling

Sequence-to-sequence Modeling

- From language modeling to seq2seq
- Vanilla encoder-decoder and its weaknesses
- Attention (remember scoring functions; no need to remember Luong/Bahdanau details)
- Training and decoding
- Issues (e.g., hallucination)
- Beam search and greedy search

Transformers

- QKV attention (important to understand well)
- Multi-head attention
- Key modules and how they fit together
- Masked attention
- Linearities and interpretability

Transfer Learning

- Understand and explain the differences between BERT, T5, GPT
- Encoder-only, Encoder-Decoder, Decoder only models
- Differences in attention, and training objectives
- Is there masking, how does masking differ across architectures?
- Conceptual differences between pertaining and fine-tuning

What is required to build GPT style models

Be familiar with various stages of development in GPT (and decoder-only LLMs), be able to explain the rationale between different model components.

- Pretraining (scaling to very large datasets)
- In-context learning (why are we giving up fine-tuning?)
- Prompting and issues with prompt formats and selection
- Instruction Tuning (what is it and why it is needed)
- RLHF, basic idea behind it (fine-tuning is back!)
- Scaling laws and why they are useful

Given a task you should be able to explain how you might implement it using prompts, and discuss the merits/disadvantages of using LLMs vs smaller models.

Predicting Linguistic Structure: POS Tagging

- Part-of-speech tags and lexical ambiguity
- The tagging task: assigning the most likely tag sequence
- Hidden Markov Models (HMMs)
  - Parameter estimation
  - Inference with the Viterbi algorithm
- Neuralized models
  - Local conditional models
  - Structured models (e.g., Markov random fields)

Predicting Linguistic Structure: Syntactic Parsing

- **Syntactic ambiguity**
  - Types and challenges
- **(Probabilistic) Context-Free Grammars**
  - Estimation for PCFGs
  - CKY algorithm for CFGs and PCFGs
  - Limitations of treebank PCFGs
  - More expressive parsing models
  - Evaluation of parsing performance

Thank You!

Thank you for attending the class!

Good luck on the exam!

Questions?