The Philosopher's guide to language modeling

Max Noichl

Utrecht University

2025-07-4

The plan

- Intro
- Bags of Words
- Topic Models
- Word-Embeddings
- BERT
- Intermezzo: OpenAlex Mapper
- GPT's
- Outro: What drives progress in philosophy?

Glossary (very sloppy)

- **Model:** A mathematical structure, implemented with computer code, intended to represent something.
- **Fitting/Training/Learning:** Jiggling around the numbers in a model, until the model predicts well. [XKCD]
- **Prediction:** Passing data to a model, and seeing what it does with it. We often do prediction on known data.
- **Supervised/unsupervised/semi-supervised:** In supervised learning we have labeled data that we learn to predict, in unsupervised learning we don't, and try to capture inherent structure. In semi-supervised learning, we make up the labels on the go.
- **Matrix:** A grid of numbers, like an excel table.

Why model language?

- Models enable computational analysis.
- Which in turn enables:
  - Large scale investigation

  - Some increase in objectivity

- General intelligence, weirdly…

Bag of words (BOW)

- The simplest of language models.
- We just count the words.

Topic models

- BOW doesn't really tell us much.
- But we can look for the hidden structures generating BOW.

Topic models (cont.)

- Problem: BOW doesn't really tell us much.
- But we can look for the hidden structures that generate a BOW.

- These are still only matrices!

Topic models (cont.)

- Problem: BOW doesn't really tell us much.
- But we can look for the hidden structures that generate a BOW.
- These are still only matrices!

- The hammer of language modeling: Not great for everything, but well known failure modes.

Word-vectors

- Problem: Topic models/BOW's are quite blunt, and don't account for semantics.
- Wittgenstein: "Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache." (PU, §43)
- Let's estimate semantics from contexts!

Word-vectors (cont.)

- Problem: Topic models/BOW's are quite blunt, and don't account for semantics.
- Wittgenstein: "Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache." (PU, §43)
- Let's estimate semantics from contexts!

- Doing math with words.

BERT

- Problem: Meaning is context sensitive.
- Solution: BERT – *Bidirectional encoder representations from transformers*
- Basic idea: Take the word-vectors, and change them depending on the other vectors that are around.
- **Attention is all you need!**

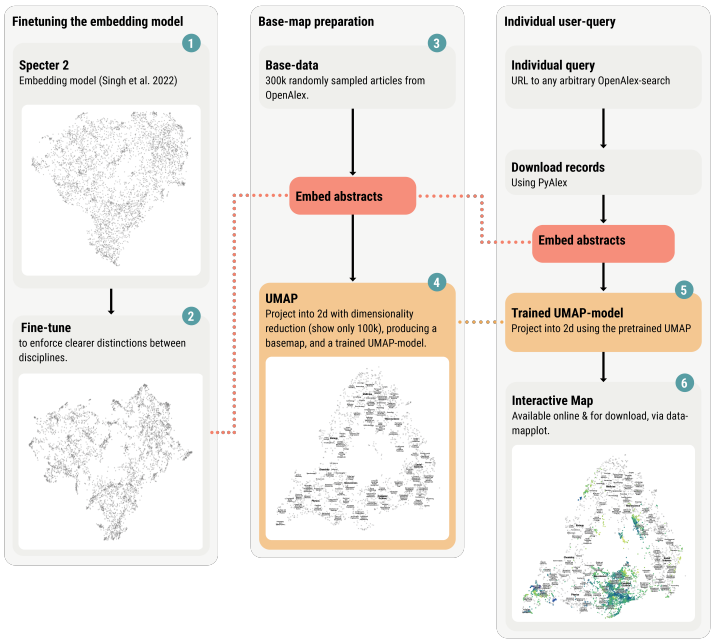
BERT (cont.)

- Problem: Meaning is context sensitive.
- Solution: BERT – *Bidirectional encoder representations from transformers*
- Basic idea: Take the wordvectors, and change them depending on the other vectors that are around.
- **Attention is all you need!**

- After contextual representation of words, just average them!

Intermezzo: OpenAlex Mapper

OpenAlex mapper



The workflow of OpenAlex Mapper, project with *Andrea Loettgers & Tarja Knuuttila @ University of Vienna* Singh et al. (2023) McInnes, Healy, and Melville (2018) De Bruin ([2022] 2023)

https://m7n-openalex-mapper.hf.space

GPTs/ 'real' LLMs

- Make BERT unidirectional (predict only ahead), make it huge, and add a sampler.
- It's just next word prediction!
- Token probabilities. (Mollick, 2024)

GPTs/ 'real' LLMs

- Make BERT unidirectional (predict only ahead), make it huge, and add a sampler.
- It's just next word prediction!

- Stochastic parrots!

GPTs/ 'real' LLMs

- Make BERT unidirectional (predict only ahead), make it huge, and add a sampler.
- It's just next word prediction!
- Stochastic parrots!

- But this is surprisingly powerful!
- Cf. Sutskever: How to predict the next word at the ending of a detective novel?
- "Know who's the culprit!"
- World models.

Intermezzo: What drives philosophical progress

Philosophical Progress

- *With Simon DeDeo (CMU, SFI)*
- Two modes of philosophical reasoning:
  - Rigorous: Propositions – Arguments – Positions
  - Fluid: Examples – Metaphors – Pictures – Intuitions
- Corpus of 23k fulltexts.
- Parse with GPT-4o into examples and positions.
- Merge: First embedd with a BERT-style model, then detect nearest neighbours, and ask gpt-4o: Are these the same?

Philosophical Progress (cont.)

- Network analysis of a bipartite network of examples and positons.
- First evidence: large examples drive positions to be successful, large positions hinder smaller ones.
- We do not just work by subdividing logical space!
- Very WIP…

Literature

De Bruin, Jonathan. (2022) 2023. "PyAlex." https://github.com/J535D165/pyalex.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." http://arxiv.org/abs/1802.03426.

Singh, Amanpreet, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. "SciRepEval: A Multi-Format Benchmark for Scientific Document Representations." November 13, 2023. http://arxiv.org/abs/2211.13308.