Exposé on Using Large Language Models for Automated Data Extraction from Scientific Literature

Felix Karg

February 15, 2023

It should be noted that everything mentioned in this document is highly uncertain, particularly scientific questions, steps and timelines, even if not otherwise mentioned.

1 Introduction

The goal of this work is to create a pipeline that derives accurate answers to given questions from a provided scientific article. For this, the article is first provided in some way: either as link to a site to download, as file-upload or through some other method to be determined later. Next, we do paragraph extraction similar to [6] to find paragraphs that describe synthesis steps. These paragraphs are then provided as context to use in answering the given questions with a Large Language Model (LLM). Answers from the extracted data will be returned directly or in a machine-readable format. From an agglomeration of such extracted data, a database may be built. This database may in further works be used in works to make predictions on MOF synthesis procedures.

2 General Topic and Motivation

While there are many articles on material synthesis, it is difficult to automatically extract important information such as reaction time, temperature, solvent, and additives in a comprehensive database. When trying to automatize the extraction of relevant data, the problem becomes one of Natural Language Processing (NLP) and proper semantic understanding of synthesis procedures.

3 Related Literature

- ChemicalTagger: A tool for semantic text-mining in chemistry [4]
  - Early demonstration that semantic data extraction works on scientific literature can work
- MOF Synthesis Prediction Enabled by Automatic Data Mining and Machine Learning [6]
  - Doing Synthesis prediction based on an automatic data extraction pipeline

- Structured information extraction from complex scientific text with fine-tuned large language models [3]
  - demonstration that LLMs can be very capable of extracting materials chemistry information for representative tasks with high accuracy

4 Scientific Questions

Benchmarking accuracy of data extraction from scientific literature using state-of-the-art LLMs and other tools: how well do various existing methods do, and (how much) can priming, fine-tuning and prompt engineering improve this.

Specifically, compare the accuracy of:

- what the model already knows: here, the article would not be in context
- what it can easily extract from the article: when provided in context, how well it can answer questions relating the content.
- particarly, without much prompt engineering or fine-tuning the model
- with prompt engineering: attempt to increase accuracy by finding good prompts for that
- after task-based fine-tuning (without context, in-context, and with or without good prompts)
  - articles included in fine-tuning
  - articles not fine-tuned on
  - experiment with fine-tuning approaches
- Models of varying sizes
  - the varying sizes of available OPT-models [10]
  - given enough time, try to use distillation [8] to compress the model parameter size

5 Intermediate Steps

1. Take exemplary / arbitrary synthesis paper
2. get paragraph classification to work
3. figure out how to pass the relevant paragraphs as LLM context
4. determine accuracy of answering the given questions based on reference database
5. first without, later with prompt engineering and fine-tuning of the model

6 Schedule

6.1 1-2 Months

- read related papers (those cited before, BERT [2], Attention is All you Need [9], GPT2 [7], GPT3 [1], Chinchilla [5] and more)
- follow Andrej Karpathy ML-Course to building a GPT-model yourself
- get OPT models of different sizes to run on cluster
- build initial pipeline to evaluate OPT models
- begin building initial dataset to

6.2 3-4 Months

- (Keep eyes out for online-mode papers, see if it could be reasonably integrated)
- finish building initial dataset to compare accuracy (before fine-tuning)
- deep probe models with (reverse) prompts, run prompt engineering experiments
- figure out ways to fine-tune, run fine-tuning experiments

6.3 5-6 Months

- Run further experiments
- keep eyes out for (new) papers on LLMs
- Mostly: Write thesis, re-run experiments (if possible / necessary)
- Deploy service, create final database with results

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S. Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. Structured information extraction from complex scientific text with fine-tuned large language models, December 2022. arXiv:2212.05238 [cond-mat].

[4] Lezan Hawizy, David M. Jessop, Nico Adams, and Peter Murray-Rust. ChemicalTagger: A tool for semantic text-mining in chemistry. *Journal of Cheminformatics*, 3(1):17, May 2011.

[5] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models, March 2022. arXiv:2203.15556 [cs].

[6] Yi Luo, Saientan Bag, Orysia Zaremba, Adrian Cierpka, Jacopo Andreo, Stefan Wuttke, Pascal Friederich, and Manuel Tsotsalas. MOF Synthesis Prediction Enabled by Automatic Data Mining and Machine Learning**. *Angewandte Chemie International Edition*, 61(19):e202200242, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.202200242.

[7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

[8] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient Knowledge Distillation for BERT Model Compression, August 2019. arXiv:1908.09355 [cs].

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, \Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[10] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open Pre-trained Transformer Language Models, June 2022. arXiv:2205.01068 [cs].