Karlsruhe Institute of Technology

AI Mat

Using Large Language Models for Automated Data Extraction from Scientific Literature

Master thesis
by

Felix Karg

Institute of Theoretical Informatics

Reviewer:	T.T.-Prof. Dr. Pascal Friederich
Second Reviewer:	Prof. Dr. Second Reviewer
Advisor:	Tobias Schlöder
Begin:	2023-03-01
Submission:	2023-09-01

Declaration of Authorship

I hereby declare that I have composed this thesis by myself and without any assistance other than the sources given in my list of works cited. This thesis has not been submitted in the past or is currently being submitted to any other examination institution. It has not been published. All direct quotes as well as indirect quotes which in phrasing or original idea have been taken from a different text (written or otherwise) have been marked as such clearly and in each single instance under a precise specification of the source.

I am aware that any false claim made here results in failing the examination.

Karlsruhe, 1st September 2023

Felix Karg

Approved as examination copy:

Karlsruhe, 1st September 2023

Contents

Abstract

Abstract goes here.

Zusammenfassung

Deutsche Zusammenfassung hier.

1. Introduction

1.1 Basics

A large amount of scientific knowledge is scattered across millions of research papers. Often, this research is not in standardized machine-readable formats, which makes it difficult or impossible to build on prior work using powerful tools to extract further knowledge.

expand on LLMs and limits here?

1.2 Motivation

Take for example the field of synthesizing Metal-Organic Frameworks (MOFs) [1]. While numerous detailed descriptions of synthesis procedures exist, they are not available in machine-readable formats, which prevents effective application of state-of-the-art techniques such as automated experimentation [2] or synthesis prediction [3]. Thus, we intend to create a pipeline for deriving machine-readable information on MOF synthesis parameters from given questions on provided scientific articles.

rewrite

2. Background

2.1 Related work

- Attention before transformers `https://jalammar.github.io/visualizing-neural-ma` `ransl` write section

Rule-Based Entity Recognition

There have long been rule-based approaches for the recognition of individual entities (e.g. Temperature). ChemTagger [4] and others [5, 6] clearly demonstrated that simple rule-based systems can sometimes extract much of the requested information. While they often achieve high precision for simple tasks, they fail in answering more complex queries, such as the relation between two entities.

Language Models

All modern language models are based on what Google introduced as the transformer architecture [7], which outperformed other available architectures with a fraction of the training cost. Based on this, Bidirectional Encoder Representation from Transformers (BERT) [8] substantially improved the state-of-the-art for all natural language processing benchmarks. BERT can be easily fine-tuned for named entity recognition in materials science [9]. Later models such as GPT2 [10] grew considerably in parameter size, as it had up to 1.5 billion parameters, up to 15x more parameters than BERT. Along with significantly increasing capability in natural language processing, these models enabled more sophisticated extraction requests. Even though automated extraction methods based on them were introduced only recently, they were already surpassed by even larger models.

Large Language Models

A continuation of increasing parameters culminated in the 175 billion parameter model GPT3 [11], the first large language model. Fine-tuning GPT3 with 100 manual and 500 partially augmented examples of data extraction created the most sophisticated pipeline for information extraction yet [12]. Most of our work will be similar to theirs. However, Chinchilla [13] and CoTR [14] demonstrated that while achieving impressive capability, such large models are substantially overparametrized and undertrained. Additionally, while the results are state-of-the-art, GPT3 is only accessible through the API of OpenAI, a for-profit company. This considerably limits access to model internals.

Our work differs from [12] by addressing these two caveats. Instead of GPT3, we use a similarly capable open-source model called OPT [15]. Self-hosting enables us to do deep introspection necessary for state-of-the-art prompt engineering and gives us the required freedom to attempt distillation [16], which addresses overparametrization. Distillation promises substantial model parameter reduction with little loss in accuracy (50x parameter reduction while keeping 95% accuracy), and has been confirmed to have similar compression characteristics for other large language models. rewrite

3. Scientific Question

The goal of this work is to use large language models to demonstrate automated extraction of unstructured text from scientific literature for the creation of a database with otherwise non-machine readable information on MOF synthesis. By doing so, we create a training pipeline that can be a) self-hosted and b) adapted to other data extraction tasks. It may be provided as a service for other research groups.

In this work, we will use OPT [15] to empirically test how much accuracy can be improved via 1) fine-tuning and 2) prompt engineering. Additionally, we intend to 3) test how accuracy and compute requirements will be affected by reduction of model size via distillation [16]. A reduction in parameters would make it considerably less compute intensive to run the final model. rewrite

4. Results

5. Conclusion

write

Appendix

Todo list

Bibliography

1. Zhou, H.-C., Long, J. R. & Yaghi, O. M. Introduction to Metal–Organic Frameworks. en. *Chemical Reviews* **112,** 673–674. ISSN: 0009-2665, 1520-6890. doi:10.1021/cr300014x. https://pubs.acs.org/doi/10.1021/cr300014x (2023-02-16) (2012-02).
2. Shi, Y., Prieto, P. L., Zepel, T., Grunert, S. & Hein, J. E. Automated Experimentation Powers Data Science in Chemistry. *Accounts of Chemical Research* **54,** 546–555. doi:10.1021/acs.accounts.0c00736 (2021).
3. Luo, Y., Bag, S., Zaremba, O., Cierpka, A., Andreo, J., Wuttke, S., Friederich, P. & Tsotsalas, M. MOF Synthesis Prediction Enabled by Automatic Data Mining and Machine Learning**. en. *Angewandte Chemie International Edition* **61,** e202200242. ISSN: 1521-3773. doi:10.1002/anie.202200242. https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202200242 (2023-02-01) (2022).
4. Hawizy, L., Jessop, D. M., Adams, N. & Murray-Rust, P. ChemicalTagger: A Tool for Semantic Text-Mining in Chemistry. *Journal of Cheminformatics* **3,** 17. ISSN: 1758-2946. doi:10.1186/1758-2946-3-17. https://doi.org/10.1186/1758-2946-3-17 (2023-02-01) (2011-05).
5. Beard, E. J., Sivaraman, G., Vazquez-Mayagoitia, A., Vishwanath, V. & Cole, J. M. Comparative Dataset of Experimental and Computational Attributes of UV/Vis Absorption Spectra. en. *Scientific Data* **6,** 307. ISSN: 2052-4463. doi:10.1038/s41597-019-0306-0. https://www.nature.com/articles/s41597-019-0306-0 (2023-02-20) (2019-12).
6. Huang, S. & Cole, J. M. A Database of Battery Materials Auto-Generated Using ChemDataExtractor. en. *Scientific Data* **7,** 260. ISSN: 2052-4463. doi:10.1038/s41597-020-00602-2. https://www.nature.com/articles/s41597-020-00602-2 (2023-02-20) (2020-08).
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. Attention Is All You Need. *Advances in neural information processing systems* **30** (2017).
8. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805.* arXiv: 1810.04805 (2018).
9. Zhao, X., Greenberg, J., An, Y. & Hu, X. T. *Fine-Tuning BERT Model for Materials Named Entity Recognition* in *2021 IEEE International Conference on Big Data (Big Data)* (2021-12), 3717–3720. doi:10.1109/BigData52589.2021.9671697.
10. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. Language Models Are Unsupervised Multitask Learners. en. *published on GitHub* (2019).

11. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. & Askell, A. Language Models Are Few-Shot Learners. *Advances in neural information processing systems* **33,** 1877–1901 (2020).

12. Dunn, A., Dagdelen, J., Walker, N., Lee, S., Rosen, A. S., Ceder, G., Persson, K. & Jain, A. Structured Information Extraction from Complex Scientific Text with Fine-Tuned Large Language Models. *arXiv:2212.05238.* doi:10.48550/arXiv.2212.05238. arXiv: 2212.05238 [cond-mat]. http://arxiv.org/abs/2212.05238 (2023-02-01) (2022-12).

13. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O. & Sifre, L. Training Compute-Optimal Large Language Models. *arXiv:2203.15556.* doi:10.48550/arXiv.2203.15556. arXiv: 2203.15556 [cs]. http://arxiv.org/abs/2203.15556 (2023-02-06) (2022-03).

14. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G. & Smola, A. Multimodal Chain-of-Thought Reasoning in Language Models. *arXiv preprint arXiv:2302.00923.* arXiv: 2302.00923 (2023).

15. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T. & Zettlemoyer, L. OPT: Open Pre-trained Transformer Language Models. *arXiv:2205.01068.* arXiv: 2205.01068 [cs]. http://arxiv.org/abs/2205.01068 (2023-02-01) (2022-06).

16. Sun, S., Cheng, Y., Gan, Z. & Liu, J. Patient Knowledge Distillation for BERT Model Compression. *arXiv:1908.09355.* arXiv: 1908.09355 [cs]. http://arxiv.org/abs/1908.09355 (2023-02-13) (2019-08).