IS20140 Final Project: Semi-automatic literature review

Samuel Dornan – 20385853

Tshitoyan et al. (2019) and Yoshitake et al. (2022) separately showed that it was possible to train a language model to infer the characteristics of certain materials using a focused corpus of text composed of abstracts from the material science domain. Raffel et al. (2020) further proved that it was possible to produce similar results by fine-tuning a large machine-learning model trained on a large corpus of generic text from across multiple domains with a second, focused corpus of domain-specific knowledge.

Given the results obtained by Tshitoyan et al. (2019) and Yoshitake et al. (2022) in the application of language models to the material science domain, and the success of the approach taken by Raffel et al. (2020) which enables end users to create powerful language models from a smaller, more accessible, more focused corpus of text, I plan to research if it is possible to apply the same methods the researchers used to a small, focused corpus of text from the social sciences domain.

The specific corpus of text I will use is obtained from the Web of Science database, and is comprised of approximately 100 abstracts of papers found using the search term “social sciences,” and the tag “Highly Cited.” The “Highly Cited” tag was used to both reduce the corpus to a more manageable size and avail of the “wisdom of the crowds” approach, which increases the likelihood that the data used to train the model is accurate, thus increasing the overall accuracy of the model and making it more likely that the model will be able to infer useful information about the domain being studied.

I may use a version of the T-5 language model originally created by Raffel et al. (2020), which was recently rearchitected by Dettmers et al. (2022) to use 8-bit matrix multiplication as opposed to 16-bit matrix multiplication, as this reduces the memory footprint of the model and enables it to be trained on publicly available computing resources.

I may add a semantic network graph to the output of the model enabling it to be represented visually, or else attach a semantic search programme to the input of the model, enabling the user to further fine-tune the model and makes it more likely that it will produce a relevant output.

Ultimately, I would like to find out if it is possible to apply the same techniques used by Tshitoyan et al. (2019) and Yoshitake et al. (2022) with great effect and success in the material science domain to the social sciences, as this would enable researchers to increase the efficiency and scale of background literature reviews, infer connections between disparate sources, disciplines, and fields and ultimately produce novel research.

References

DETTMERS, T., LEWIS, M., BELKADA, Y. & ZETTLEMOYER, L. 2022. LLM. int8 (): 8-bit Matrix Multiplication for Transformers at Scale. *arXiv preprint arXiv:2208.07339*.

RAFFEL, C., SHAZEER, N., ROBERTS, A., LEE, K., NARANG, S., MATENA, M., ZHOU, Y., LI, W. & LIU, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.,* 21**,** 1-67.

TSHITOYAN, V., DAGDELEN, J., WESTON, L., DUNN, A., RONG, Z., KONONOVA, O., PERSSON, K. A., CEDER, G. & JAIN, A. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature,* 571**,** 95-98.

YOSHITAKE, M., SATO, F., KAWANO, H. & TERAOKA, H. 2022. MaterialBERT for natural language processing of materials science texts. *Science and Technology of Advanced Materials: Methods,* 2**,** 372-380.