LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Embedding arithmetic in Transformer embedding matrices?**

Supervisor: Sebastian Gerstner

Examiner: Hinrich Schütze

Level: BSc / MSc

Summary: Mikolov, Yih, and Zweig 2013 discovered that their word embeddings enable *embedding arithmetic*: For example, in vector representation, $\texttt{king} \approx \texttt{queen} - \texttt{woman} + \texttt{man}$. Since then, Transformer-based language models have appeared. They also include embeddings (for tokens), but these are trained in a quite different manner. You will research to what extent embedding arithmetic is also possible in the embedding matrix of a Transformer model.

BSc:

Try to replicate Mikolov, Yih, and Zweig 2013 with a Transformer embedding / unembedding matrix (e.g. that of GPT2). Where are the results similar, where are they different?

MSc (additionally):

There will be linguistic or semantic relations that cannot be described in terms of vector addition so well - at least this will be the case of many-to-many relations. Can you find another mathematical description for these (or some of these)? Ideas: a linear or affine map à la Hernandez et al. 2023, or anything you find in Wang et al. 2017.

An additional, independent question: in more recent LLMs embedding and unembedding matrices are different - analyse both of them and compare your results.

Finally, try to give a theoretical explanation for your results.

Requirements: Programming, linear algebra. It's a good thing if you have previously heard of embedding arithmetic and/or know about Transformer-based language models.

References:

- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic Regularities in Continuous Space Word Representations". In: *North American Chapter of the Association for Computational Linguistics*. URL: `https://aclanthology.org/N13-1090.pdf`
- Evan Hernandez et al. (2023). "Linearity of Relation Decoding in Transformer Language Models". In: *ArXiv* abs/2308.09124. URL: `https://lre.baulab.info`
  Evidence that linear or affine maps can describe relations well. They look at the whole computation of a Transformer, you care only about embeddings.
- Quan Wang et al. (2017). "Knowledge Graph Embedding: A Survey of Approaches and Applications". In: *IEEE Transactions on Knowledge and Data Engineering* 29, pp. 2724–2743. URL: `https://api.semanticscholar.org/CorpusID:19135805`
  Further ideas about how relations could be represented in embedding space. You need only sections 3.1 and 3.2. Remember that your entities (words / tokens) are represented as deterministic vectors in a real vector space, so not all models qualify.