Understanding LLMs
The Role of Word Embeddings

How does an LLM Understand a Word?

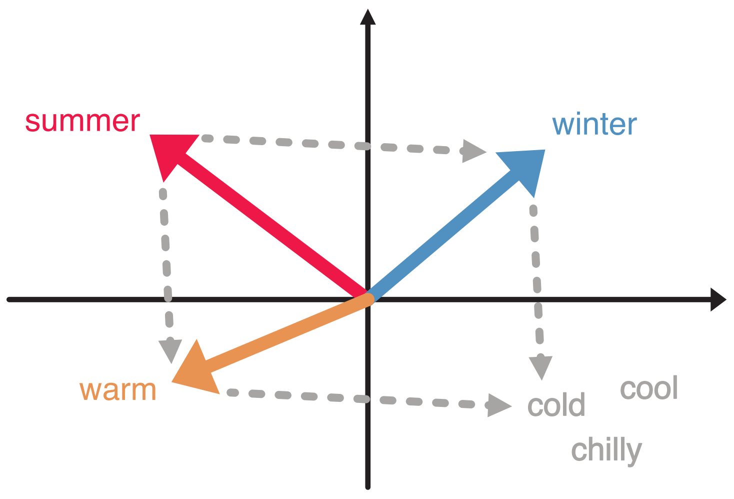
In the realm of Natural Language Processing (NLP), Large Language Models (LLM) are gaining more popularity for imitating human language. These advanced computer frameworks use deep learning techniques to efficiently analyze, understand and generate language at a near-human level. At the core of these models is the concept of word embeddings. (Naveed et al., 2024)

Word Embeddings

Word embeddings are dense vector representations of words in a continuous space where semantically similar words are positioned closely together. These embeddings capture the nuanced relationships between words, allowing models to understand and generate human-like text. Unlike traditional one-hot encoding, which represents words as sparse vectors with high dimensionality, word embeddings transform words into lower-dimensional vectors that encode semantic meaning. (Fares et al., 2017)

How LLMs use Word Embeddings

LLMs use word embeddings to interpret words based on their context within sentences or larger bodies of text. This contextual understanding is achieved through algorithms like *word2vec*, which train on vast amounts of text data to learn the relationships between words. By processing these embeddings, LLMs can perform a variety of language tasks with remarkable accuracy, from translation and sentiment analysis to more advanced applications like automated customer service and content generation. (Fares et al., 2017)



video with visualizations for word embedding vectors

`https://www.youtube.com/watch?v=wjZofJX0v4M`

Semantic Calculator with WebVectors for word relations

`http://vectors.nlpl.eu/explore/embeddings/en/calculator/`

How are Different Words Connected?

Each word is represented by a vector. This allows for the use of vector arithmetic to uncover word relationships and analogies.

Semantic Calculator with WebVectors

This tool utilizes pre-trained word embeddings to demonstrate the power of vector calculations in understanding and manipulating word relationships. To try the Semantic Calculator, scan the blue QR-Code. It is possible to explore word relations through vector operations. For instance, by adding and subtracting word vectors, one can find analogies such as

```
king - man + woman = queen
```

Experiment locally

It is possible to try different word embeddings locally with for example Python. A variety of different datasets for the local example are available at
`https://github.com/piskvorky/gensim-data`

Want to try it out locally? (Gensim, 2022)
Installation procedure and further information at
`https://gitlab.tugraz.at/ed-tech/2024/role-of-word-embeddings`

Python 3 example

```
$ python3
Python 3.9.12
> import gensim.downloader as api
> dataset = api.load('glove-wiki-gigaword-300')
> dataset.most_similar_cosmul(positive=["warm", "winter"], negative=["summer"])
cold
> dataset.most_similar_cosmul(positive=["boat", "road"], negative=["sea"])
bicycle
```

Why is the Training Material Important?

The quality and size of the training material are important for the performance of language models. For natural language processing, the datasets used to train these models serve as the basic structure that shapes their ability to understand and generate human language. Larger datasets generally provide a broader spectrum of linguistic contexts, idiomatic expressions, and domain-specific terminologies, which strengthen the model's versatility and accuracy.

To prove that the quality and size of the training material are important for the performance of language models, think about the following analogy task: a car is related to a road like a boat is related to **?**

Let's try this with the Semantic Calculator. When using the English Wikipedia model with 3 billion tokens, the word "roads" is selected, while with the English Gigaword model with 4.8 billion tokens, the word "river" is selected. What's your opinion? Which answer suits better?

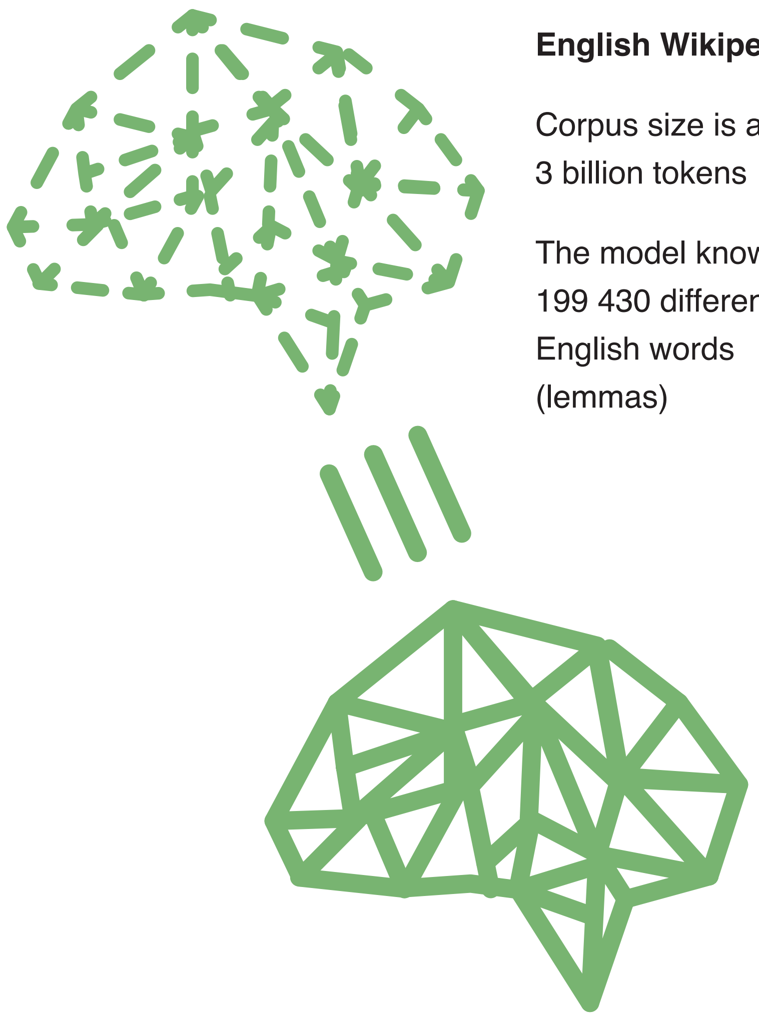
```
car - road + boat = roads @ English Wikipedia

car - road + boat = river @ English Gigaword
```

Invent and Test an Analogy Task Yourself!
On the Semantic Calculator website you can select different models and test their performance yourself.

Different Model Sizes

from `http://vectors.nlpl.eu/explore/embeddings/en/models/`



English Wikipedia

Corpus size is about 3 billion tokens

The model knows 199 430 different English words (lemmas)

English Gigaword

Corpus size is about 4.8 billion tokens

The model knows 297 790 different English words (lemmas)

Sources

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). *A Comprehensive Overview of Large Language Models (Version 9).* arXiv. `https://doi.org/10.48550/ARXIV.2307.06435`

Fares, Murhaf; Kutuzov, Andrei; Oepen, Stephan & Velldal, Erik (2017). *Word vectors, reuse, and replicability: Towards a community repository of large-text resources*, In Jörg Tiedemann (ed.), Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017. Linköping University Electronic Press. ISBN 978-91-7685-601-7 `https://www.duo.uio.no/bitstream/handle/10852/65205/ecp17131037.pdf`

Gensim (2022). *API Reference.* https://radimrehurek.com/gensim/apiref.html (last accessed 2024-05-27)

Poster online avialable at https://doi.org/10.5281/zenodo.11348467