SURV 622/SURVMETH 622: TEXT ANALYSIS AND LARGE LANGUAGE MODELS (LLMs)

Mao Li

Michigan Program in Survey and Data Science (MPSDS)
Michigan Institute for Computational Discovery and Engineering (MICDE)

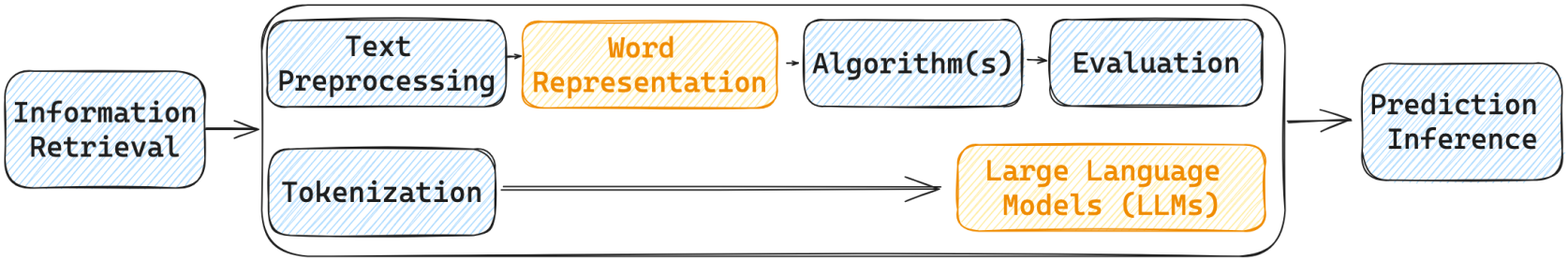
March 24, 2025

Outline

What is Quantitative Text Analysis?

- Broad definition: Process of converting words to numbers
- Can be done manually, but for large amounts of text automated methods will be essential
- Can involve various steps for processing and analysis

Text Analysis Pipeline



A long history exists of human efforts to enable machines to communicate with humans.....

Outline

Bag of Words Model

- In this model, a text (such as a sentence or a document) is represented as the bag (collection) of its words, disregarding grammar and even word order but keeping frequency.
- Document-term matrix:
  - Document: a basic unit of textual data.
  - Term: a piece of a document, usually an individual word.
- Consider this document: "The quick brown fox jumps over the lazy dog"
  - Can split into 9 words: [The] [quick] [brown]. . . [dog]
  - Can split into 8 bigrams: [The quick] [quick brown] [brown fox]. . . [lazy dog]

Bag of Words Model (cont'd)

Term Frequency(TF)

- Just Count and divide: how often the word appears in the corpus.

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

where $f_{t,d}$ is the raw count of a term in a document.

Term Frequency-Inverse Document Frequency (TF-IDF)

- intended to measure how important a word is to a document

$$idf(t,D) = log \frac{N}{|\{d \in D : t \in d\}|}$$

$$tfidf = tf(t,d) \times idf(t,D)$$

where N is the total number of documents in the corpus, and $|\{d \in D : t \in d\}|$ number of documents where the term t appears.

Demo

Google Colab

Limitations (Bag Of Words)

Simplifying Assumptions

- Word order does not matter
  - In reality, the position of words relative to each other is often critical in conveying intention behind a sentence. E.g., "John is taller than Mary." $\neq$ "Mary is taller than John."
- Each word has one and only one meaning
  - In reality, words can have different meanings based on their appearance alongside other words. E.g., "running" for exercise vs. "running" for office.

Problems

- Vector space is high-dimension and sparse.
- Each word in the vector space is orthogonal to each other.

Representation of Words

Nets are for fish; Once you get the fish, you can forget the net. Words are for meaning; Once you get the meaning, you can forget the words

— Zhuangzi, Chapter: External Things (Wai Wu)

Distributed Representation

In: *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. Amherst, Mass. 1986, pages 1-12.
Erlbaum, NJ.

LEARNING DISTRIBUTED REPRESENTATIONS OF CONCEPTS

Geoffrey E. Hinton
Computer Science Department
Carnegie-Mellon University

Word2Vec

Goal

- Computing continuous vector representations of words and learning word associations from a large corpus of text.

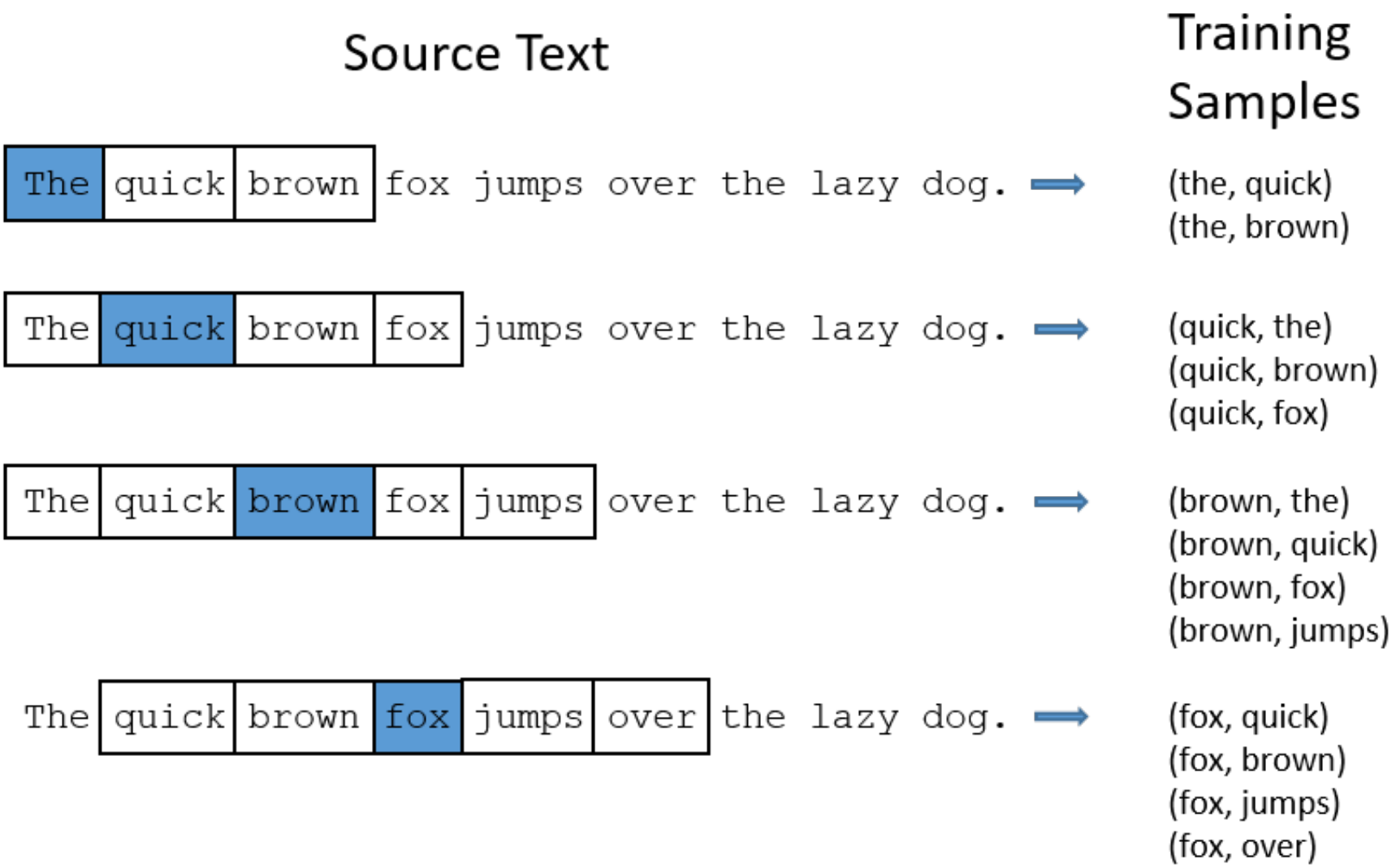
Model Architectures

- Continuous Bag-of-Words Model (CBOW)
- Continuous Skip-gram Model

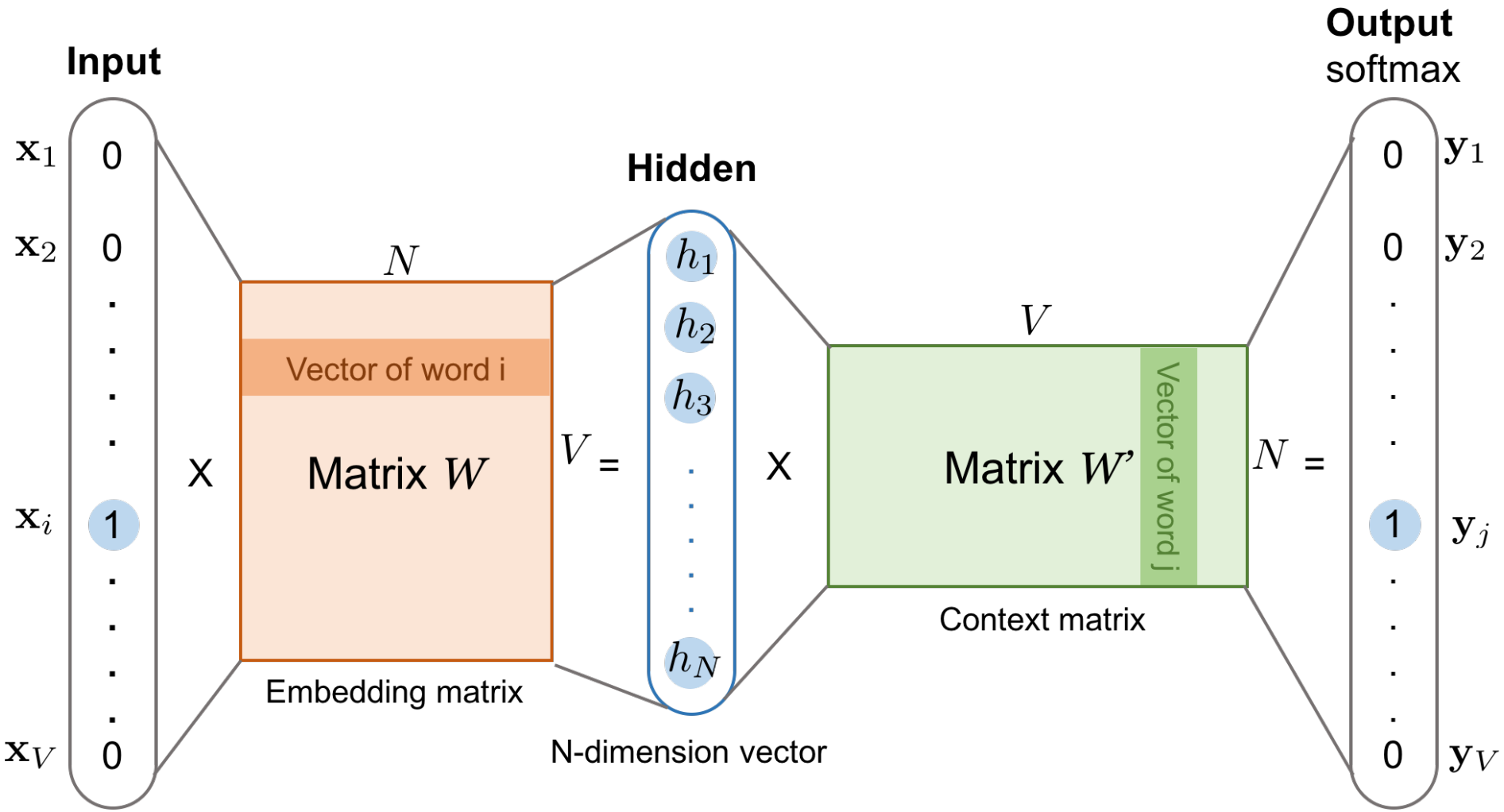
Advanced approaches

- Subsampling Frequent Words
- Negative Sampling

Contiuous Skip-gram Model



Contiuous Skip-gram Model



Subsampling

- Frequent words occur millions of times (e.g., "in", "the", and "a") while provide less information value than rare words.
- Each word w_i in the training set is discarded with probability computed by:

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$$

Negative sampling

- Distinguish the target word from "negative samples".

While NCE can be shown to approximately maximize the log probability of the softmax, the Skip-gram model is only concerned with learning high-quality vector representations, so we are free to simplify NCE as long as the vector representations retain their quality. We define Negative sampling (NEG) by the objective

Group Exercise (10 minutes)

Try the Word2Vec by yourself:

- Arithmetic operations on vectors (plus, minus, dot product, etc.)
- Calculate the similarity among word pairs you are interested in.
- Discover potential implicit stereotypes in the Word2Vec model.

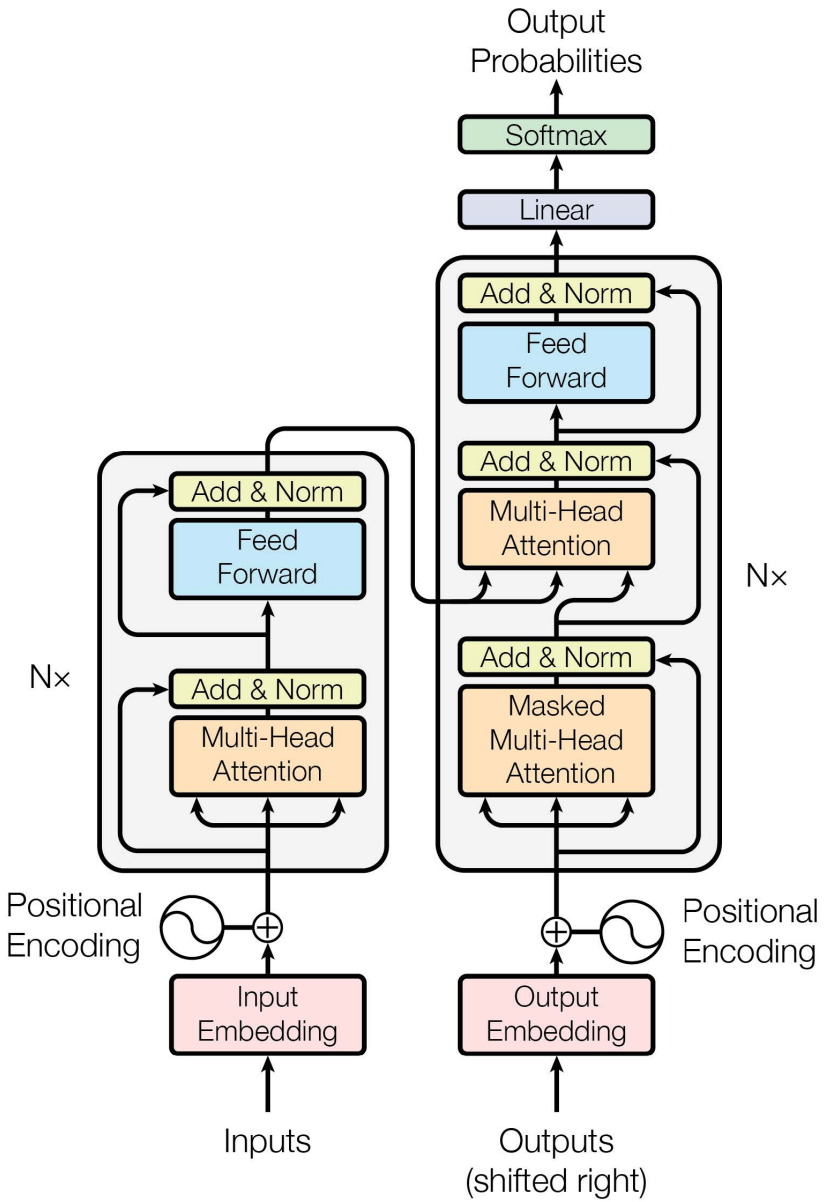
Limitations (Word2Vec)

- The window size is fixed, and each in-context word has the same "significance".
- Still, each word has one and only one meaning.

Transformer Architecture

Attention is All You Need!

- An architecture that aims to solve sequence-to-sequence tasks while easily handling long-distance dependencies.
- The overall architecture can be broken down into two different parts: Encoder and Decoder.
  - Encoder-only models: BERT, RoBERTa, etc.
  - Decoder-only models: GPTs, Llama, etc.



Transformer Architecture (cont'd)

- Self-Attention Mechanism:

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

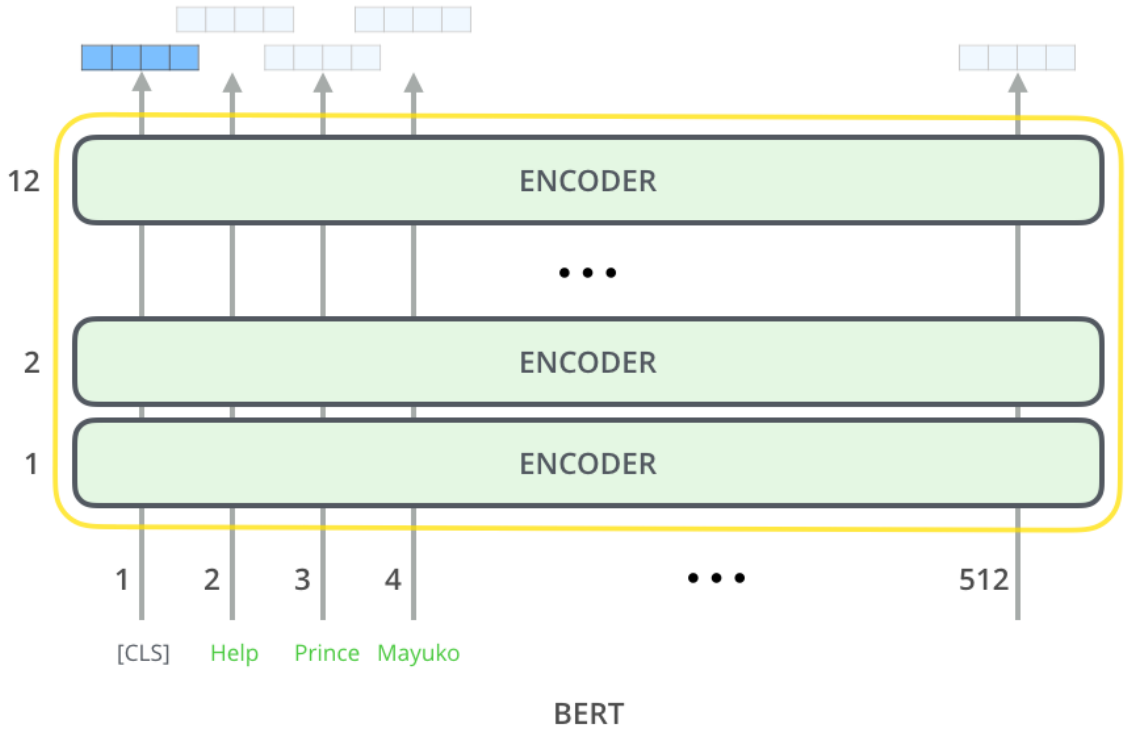
  Where Q represents the matrix of queries, K represents the matrix of keys, V represents the matrix of values, and d_k is the dimension of the keys and queries used for scaling the dot products.

- After determining the "similarity" between vectors representing word meanings, the mechanism prioritizes words with closer associations, giving them higher significance and, thus, greater activation levels. While in backpropagation stage, if a particular activation contributes positively to the prediction outcomes, the linkage between the relevant words is reinforced and vice versa. In this way, the model learns the intricate relationship between words.

BERT: Bidirectional Encoder Representations from Transformers

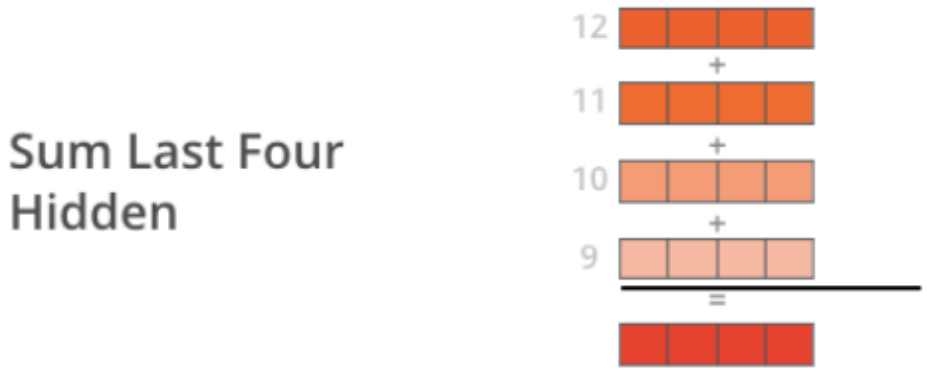
- **Advancement in Embedding: BERT** enhances the concept of language model-driven word embedding, building on ELMo's foundation.
- **Bidirectional: BERT** uniquely processes sentences in both directions, ensuring comprehensive context understanding.
- **Encoder:** specializes in encoding input sequences to capture nuanced text features effectively.
- **Pretraining: BERT** sets a new standard by pretraining on a vast corpus of text, leveraging unlabelled data to understand language nuances before fine-tuning for specific tasks.

BERT Architecture



BERT for Word Representation

- Each encoder layer can represent a word vector!
- The two common practices are
  1. pooling the last four encoder layers from **BERT**.



  2. concatenating the last four encoder layers.

Demo

Google Colab

Outline

GPT: Generative Pretrained Transformer

Background

- Paradigm introduced by **BERT**
  - Pretrain + Finetune: The language model will undergo an initial pretraining phase on extensive text corpora, acquiring a broad understanding of language patterns.
  - For specific applications like text classification or question answering, users finetune the model on task-specific datasets, optimizing BERT for particular requirements.
- How can we achieve AGI (Artificial General Intelligence)?
  - Achieving AGI remains a challenge within this model training paradigms.
  - Models like BERT require finetuning with task-specific data, indicating that building a truly "general" model capable of understanding and performing any intellectual task that a human being can is still beyond our reach.

GPT: Generative Pretrained Transformer

Language Models are Few-Shot Learners!

- OpenAI argues that nearly any NLP task can be reformulated as a next token prediction problem, a groundbreaking approach allowing for various applications.
- Large Language demonstrates remarkable few-shot learning capabilities, understanding and performing tasks with minimal examples. This contrasts with traditional models requiring extensive finetuning data.

Understand Data Flow in Large Language Models

- **Tokenization**: Initially, the input text is broken down into manageable segments or tokens, a process that prepares the data for detailed analysis by the model.
- **Layered Processing**: These tokens are then sequentially processed through the multiple layers of the Large Language Model. Each layer builds on the previous one to enhance the model's understanding and interpretation of the text.

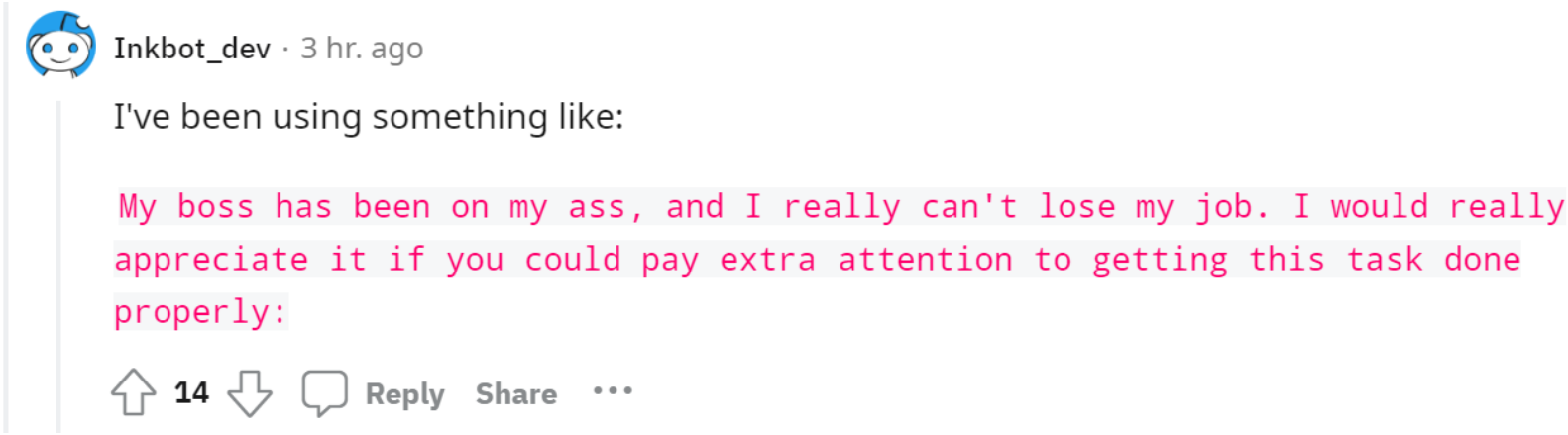
All about Prompt

In the era of Large Language Models (LLMs), Prompt Engineering is increasingly recognized as a critical skill:

- **Zero/Few-Shot Learning**: This method involves designing prompts that enable the model to undertake new tasks without explicit prior training, using either a task description or a small number of examples within the prompt itself.
- **Chain of Thought (CoT) Prompting**: Tailored for complex reasoning tasks, CoT prompting instructs the model through a logical sequence of steps. This structured guidance assists the model in addressing challenges that demand complicated inferential reasoning.

Taking a Small Detour...

How to enhance the prompt:



Inkbot_dev · 3 hr. ago

I've been using something like:

```
My boss has been on my ass, and I really can't lose my job. I would really appreciate it if you could pay extra attention to getting this task done properly:
```

14 Reply Share ···

Large Language Models Understand and Can Be Enhanced by Emotional Stimuli

Cheng Li[1], Jindong Wang[2]*, Yixuan Zhang[3], Kaijie Zhu[2], Wenxin Hou[2], Jianxun Lian[2], Fang Luo[4], Qiang Yang[5], Xing Xie[2]
[1]Institute of Software, CAS [2]Microsoft [3]William&Mary
[4]Department of Psychology, Beijing Normal University [5]HKUST

Outline

Why Stance?

- Long tradition in the social sciences of analyzing unconstrained language produced by, for example
  - research participants, in open survey responses
  - members of the public, e.g., in social media posts
  - journalists, e.g., reporting or opinion makers, e.g., editorials, speeches
  - Until recently, unconstrained text was manually coded, e.g., through crowdsourcing such as MTurk
- For large corpora such as social media posts, less content-related metrics have been introduced because that's all the technology could handle:
  - Volume: only shows the popularity of topics.
  - Sentiment: the mood of the public, but sometimes the sentiment is unrelated to opinion.
- Stance: the attitude, position, or viewpoint that an individual adopts with respect to a specific topic, issue, or debate.

Sentiment VS. Stance

Stance can be orthogonal (unrelated) with sentiment

- For instance, the debate surrounding the inclusion of a citizenship question in the 2020 Census has generated a considerable amount of controversial discourse on Twitter.

Examples of Tweets Opposing Citizenship Question in 2020 Census

Tweet	Sentiment	Stance
I'm just glad Trump didn't get to include the citizenship question he wanted.	Positive	Oppose
Trump will damage and start more trouble around the country b/4 the election. He now is trying to screw up the census with an Executive order that will only count citizens, when counting total population is how it's been done for however long there has been a census.	Negative	Oppose

Align Survey with Social Media Posts Using Stance

- O'Coonnior et al. (2010) demonstrated a correlation between ICS and daily sentiment (OpinionFinder) of jobs tweets is 0.65. However, the results of Conrad et al. (2021) showed that the correlation scores between ICS and OpinionFinder highly depend on how sentiment is calculated.
  - This suggests that relying solely on sentiment analysis may not provide a consistent measure for aligning with survey estimates.
- Conrad et al. (2023) further proposed uncovering alignment between survey responses and social media by detecting the stance expressed in posts.
  - By focusing on stance, they discovered meaningful correlations between specific Census Tracking Survey Questions and social media posts about the Census, revealing alignments that sentiment analysis alone had obscured.

Summary

- **Optimizing Word Representation:** Selecting the appropriate representation for words or texts is crucial in text analysis. Researchers must identify and employ the most effective representation strategy to enhance decision-making.
- **Purpose and Approach to Text Analysis:** The primary objective of text analysis is to derive meaningful insights from unstructured text data. However, the focus should not solely be on the methodology but rather be guided by the specific research questions at hand. Adopting a question-driven approach ensures that the analysis is both relevant and purposeful.
- **The Role of Large Language Models (LLMs):** LLMs significantly reduce the analytical load on researchers by shifting the emphasis away from data towards the formulation of the prompt. In this new paradigm, the critical task is to pose the most appropriate prompt/instruction to LLMs, leveraging their capabilities to uncover deep insights from textual data.