PLIN021 SÉMANTICKÁ ANALÝZA V PRAXI

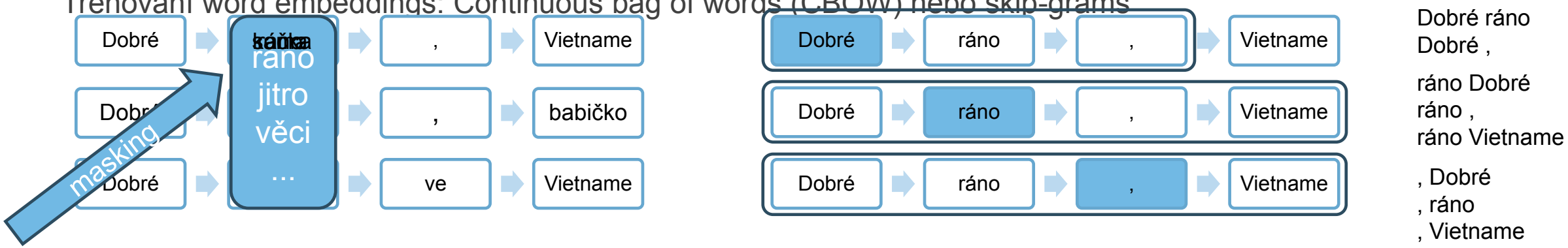
ZUZANA NEVĚŘILOVÁ

2020/21

VEKTOROVÁ REPREZENTACE SLOV

- Slovo (token) je popsáno vektorem: sparse vector (hodně nul nebo stejných čísel) a dense vector (různá čísla).
- Vektor **reprezentuje určité aspekty** významu:
  - One hot – různá slova mají různé vektory
  - Dense vector (word embeddings) – slova, která se nacházejí v podobné společnosti, mají podobné vektory

Trénování word embeddings: Continuous bag of words (CBOW) nebo skip-grams



WORD EMBEDDINGS

- Slovo (token) je popsáno vektorem, nejčastější dimenze je 50-300.
- Slova z podobných kontextů mají podobné vektory (svírající malý úhel, na směru šipky záleží víc než na délce).
- Pokud vektor kóduje významy slova, kóduje všechny významy slova, které se vyskytují v korpusu.
- Vektor **nekóduje funkci** slova ve větě.

Pilný syn líného otce. Líný syn pilného otce.

- Trénování pomocí CBOW nebo skip-gram má limit ve velikosti okna. Co když je potřebný kontext delší?

Žádáme výměnu koberce pro našeho předsedu, který je sešlý a k ničemu se nehodí.

- Pro kontext jsou důležitá slova v různé vzdálenosti.

TRANSFORMERY

- Jak dlouhý kontext se bere v potaz při trénování sítě? Max. limit bývá 512-1024 tokenů, ne všechna slova jsou důležitá.
- Transformer jen některým slovům věnuje větší **pozornost** (attention).
  - Předchozí modely věnují pozornost posledně viděnému vstupu nebo n posledně viděným vstupům.
  - Rekurentní neuronové sítě (RNN) zpracovávají sekvence dat (informace proudí dopředu a dozadu v dané sekvenci).
- Self-attention – jak důležitá jsou předchozí slova pro predikci tohoto slova?
  - Pozornost (attention) je vážený součet **všech** předchozích stavů.
  - „Předchozí“ se neuvažují v sekvenci, ale jako množina (nezáleží na pořadí). Vzdálenost mezi všemi vstupy je stejná.
- Transformer je sekvenční model, který využívá self-attention pro propagaci informací.
  - Maskování (masking) změní self-attention na sekvenční operaci (causal transformer block).
  - Kódování pozice tokenů (position embeddings, position encodings, relative positions)

GENERATIVNÍ MODELY

- Pravděpodobnostní model
  - Diskriminativní model – podmíněná pravděpodobnost
    Jaká je pravděpodobnost, že pozorované zvíře je **kráva** (cíl), když má pozorování tyto prvky (**býložravec**, **velký**)? P(Y | X=x)
  - Generativní model – sdružená/simultánní distribuční funkce (joint probability distribution) pozorování a cíle P(X, Y) nebo P(X | Y=y)
- Generativní AI
  - Zakóduje vstupy, naučí se z nich vzory, z nichž generuje výstupy s podobnou charakteristikou, jako měla vstupní data
  - Text-to-text: GPT, Bard, LLaMA
  - Text-to-image: DALL-E, Midjourney
- Modality: text, kód, obraz, audio, video, chemické vazby, pohyb, ...

VELKÉ JAZYKOVÉ MODELY (LARGE LANGUAGE MODELS, LLMS)

- Velké (175B – miliard – parametrů – což je velikost GPT3)
- Jazykové (přirozený jazyk)
- Model (statistický popis sekvencí slov)

Jazykový model – natrénovaný z korpusů (statistika – pravděpodobnosti slov)

Parametry modelu

- Nesouvisí s počtem slov, ale s architekturou neuronové sítě
  - Jsou to váhy (desetinná čísla) v neuronové síti
- Některé vrstvy (layers) zachycují jednodušší aspekty slov (např. slovní druh), jiné mohou kódovat komplexní vzory

JSOU LLMS TYPEM GENERATIVNÍ AI? NE TAK ÚPLNĚ.

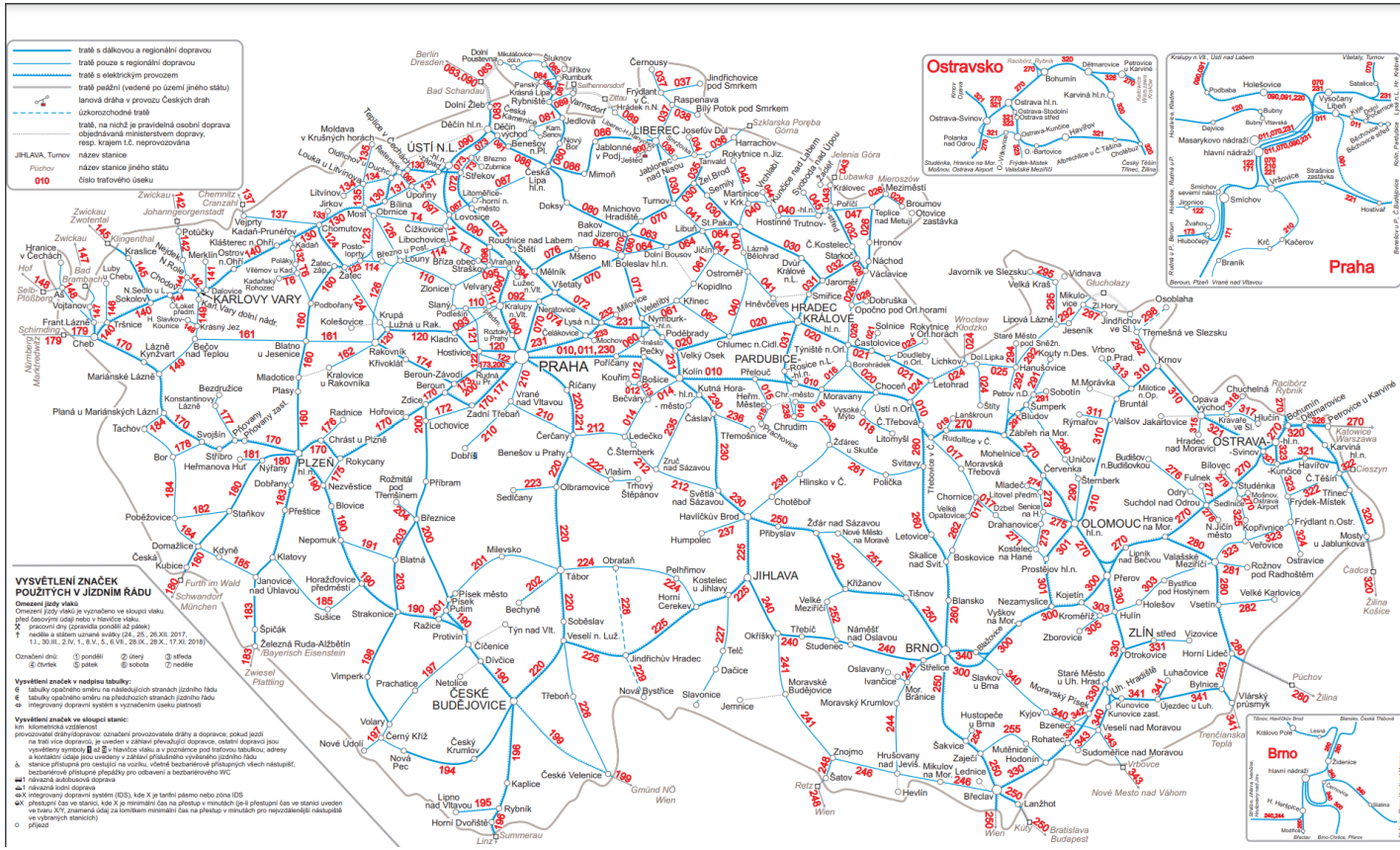
- ChatGPT je generativní systém, který využívá LLM

GPT3, převzato z (Verspoor, 2023)

- LLMs se používají pro klasifikační úlohy (analýza sentimentu)
- LLMs se používají pro seq2seq úlohy (strojový překlad)
- LLMs se používají pro zakódování textu pro následné predikce (downstream tasks)
- GPT = generative + pre-trained + transformer

GENEROVÁNÍ TEXTU

- Jaký je nejpravděpodobnější další token, když je dána sekvence tokenů $t_1, \ldots, t_n$?
  - Hladové prohledávání (Greedy search)
- Jaké jsou nejslibnější další tokeny?
  - Paprskové prohledávání (Beam search)
- Jaké jsou *k* nejpravděpodobnější další tokeny, jejichž kumulativní pravděpodobnost *p* je vyšší než práh?
  - Top-k a top-p vzorkování (Sampling top-k and top-p)



LITERATURA

- McCormick, C. (2016, April 19). *Word2Vec Tutorial - The Skip-Gram Model*. Retrieved from http://www.mccormickml.com. https://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
- Peter Bloem: Transformers from scratch. 2019. Vrije Universiteit Amsterdam. https://peterbloem.nl/blog/transformers
- Karin Verspoor: Large Language Models Are Not (Necessarily) Generative Ai. Open Data Science. 2023. https://www.youtube.com/watch?v=vhrMCVdJbU4