Témata závěrečných prací a ročníkových projektů / Topics for individual software projects, bachelor and master theses

Pokud vás nějaký z nápadů zaujal, napište mi email na libovicky@ufal.mff.cuni.cz

If you're interested in any of these topics, please drop me a line at libovicky@ufal.mff.cuni.cz

Měření kulturních hodnot ve velkých jazykových modelech napříč jazyky / Measuring cultural values in large language models across languages

O jazykových modelech je známo, že zachycují kulturní vzorce (a s nimi často různé více či méně žádoucí stereotypy) z trénovacích dat. Existující metody vyhodnocování toho, jaké hodnoty jsou v modelu zastoupeny většinou nechají jazykové modely vyplňovat dotazníky jako by je vyplňovali lidé. Za tím stojí předpoklad, že jazykový model se chová jako singulární agent. Výzkumy ale ukazují, že jazykové modely se chovají spíše jako ensemble agentů. V tomto pohledu jsme ale získali odpověď pouze od jednoho z agentů. Správná metodologie by tedy měla zahrnovat spíš samplování různých odpovědí věčetně různých zdůvodnění. Cílem práce bude vyvinout spolehlivější evaluační metodologii, která poskytne citlivější odhad jaké hodnoty jsou v jaké míře zastoupené v různých modelech.

Language models are known to capture cultural patterns (and with them often various more or less desirable stereotypes) from training data. Existing methods for evaluating what values are represented in a model usually let language models fill in questionnaires as if humans filled them in. Behind this is the assumption that the language model behaves as a singular agent. However, research shows that language models behave more like ensemble agents. However, we only got an answer from a single agent. A proper methodology should, therefore, involve sampling different answers, including different justifications. The goal of this work will be to

develop a more reliable evaluation methodology that provides a more sensitive estimate of what values are represented to what extent in different models.

Co se naučíte: pracovat s velkými jazykovými modely napříč jazyky, evaluovat komplexní experimenty

Skills you'll acquire: work with large language models across languages, evaluation of complex experiments

Výpočetně efektivní vícejazyčné větné embeddingy / Highly efficient sentence embeddings

Vektorové reprezentace vět, také nazývané větné embeddingy, se dnes používají pro přenos jednodušší klasifikačních úloh mezi jazyky (máme trénovací data pouze v jednom jazyce, ale potřebuje, aby model fungoval ve více jazycích) a pro vyhledávání paralelních trénovacích vět pro strojový překlad. Větný embedding je typicky výstupem neuronové sítě typu Transformer a výpočet tak embeddingu trvá relativně dlouho. Efektivní výpočet navíc vyžaduje počítání na GPU. Cílem práce bude vyvinout metody, jak získávat větné embeddingy efektivněji pomocí jednodušší neuronových sítí, které by se naučily pomocí knowledge distillation.

Sentence vector representations, also called sentence embeddings, are nowadays used to transfer simpler classification tasks between languages (we have training data in only one language but need the model to work in multiple languages) and to find parallel training sentences for machine translation. Sentence embeddings are typically the output of a Transformer neural network, so computing such embeddings takes a relatively long time. Moreover, efficient computation requires GPU. The goal of this work will be to develop methods to obtain sentence embeddings more efficiently by using simpler neural networks that learn by knowledge distillation.

Co se naučíte: technické details architektury Transformer, trénování modelů pomocí knowledge distilation, práce s vícejazyčnými data

Skills you'll acquire: technical details of the Transformer architecture, training using knowledge distillation, work with multilingual data

Vícejazyčná klasifikace kulturně nebo politicky relevantních témat v textech / Multilingual classification of culturally or politically relevant topics in texts

Vícejazyčné modely umožňují přenést schopnost řešit úlohy z jednoho jazyka do druhého, obvykle s určitou ztrátou kvality. Při tom obvykle předpokládáme určitou podobnost napříč

jazyky a kulturami. Jazykové prostředky, které jsou směrodatné v některých společensky a kulturně podmíněných úlohách, jako je detekce nenávistných projevů nebo detekce politicky nevyvážených zpráv, se mohou v různých jazycích lišit, což může naopak negativně ovlivnit kvalitu výstupů modelů napříč jazyky. V tomto projektu bychom se zaměřili na jednu takovou úlohu, analyzovali bychom rozdíly ve výkonnosti napříč jazyky a experimentovali bychom s možným zlepšením přenosu mezi jazyky.

Multilingual language models allow the transfer of the ability to solve tasks from one language into another, typically with some performance loss. Here, the underlying assumption is similar linguistic means across languages and cultures. However, the linguistic means that are indicative in some culture-sensitive tasks, such as hate speech detection or hyper-partisan news detection, might differ across languages, which can, in turn, negatively affect cross-lingual performance. In this project, we would focus on one such task, analyze the performance differences across languages, and experiment with potential improvements in cross-lingual transfer.

Co se naučíte: technické detaily fungování vícejazyčných jazykových modelů, dotrénování jazykových modelů, práce s vícejazyčnými daty

Skills you'll acquire: technical details of multilingual language models, finetuning language models, working with multilingual data

Analýza podslovních embeddingů ve velkých jazykových modelech napříč jazyky / Analysis of subword embeddings in large language models

Podslovní embeddingy jsou úplně první vrstvou ve velkých jazykových modelech. Tokeny, které reprezentují, jsou výsledkem jednoduché statistické heuristiky: mnoho z nich odpovídá slovům nebo fragmentům slov v jazycích, které jsou dominatně zastoupeny v trénovacích datech. Jiné jsou sdílené napříč jazyky, někdy mají v jazycích stejný, jindy odlišný význam. Velké jazyky mají dlouhé tokeny specifické pro daný jazyk, zatímco texty z menších jazyky jsou segmentovány na dlouhé posloupnosti tokenů. To vede k hypotézám, že tokenizace a embeddingy v první vrstvě mohou hrát zásadní roli v rozdílech ve výkonnosti LLM v různých jazycích. Analýza toho, jaké informace o slovech a jazycích nesou podslovní embeddingy, by nám mohla pomoci pochopit rozdíly mezi jazyky ve velkých jazykových modelech.

Subword embeddings are the very first layer in large language models. The tokens they represent are a result of a simple statistical heuristic: many of them correspond to words or word fragments in languages that are dominant in the training data. Other tokens are shared across languages, sometimes having the same and sometimes a different meaning in the respective languages. Large languages have long language-specific tokens, whereas texts in small languages are segmented into long sequences of tokens heavily shared across languages. This leads to speculations that the tokenization and first-layer embeddings might play a crucial role in

the differences in LLM performance across languages. Analysis of what information about the words and language the subword embeddings carry might help us understand the differences between languages in large language models.

Co se naučíte: technické detaily fungování LLMs, metody pro interpretabilitu embeddingů, práce s vícejazyčnými korpusy

Skills you'll acquire: technical details of LLMs, interpretability methods for embeddings, work with multilingual corpora

Dekódování v neuronovém strojovém překladu s explicitním odhadem délky věty / Neural machine translation decoding with explicit length estimation

Současné modely pro neuronový strojový překlad modelují pravděpodobností distribuci na cílovými větami různých délek, ale neříkají, jak z takového modelu získat nejpravděpodobnější větu v cílovém jazyce. V současnosti se používá heuristické prohledávací algoritmy, které se zároveň snaží kompenzovat, že počet vět určité délky roste exponenciálně s délkou věty. Kdybychom uměli spolehlivě odhadnout délku cílové věty, celý proces dekódování by to mohlo zjednodušit.

Current models for neural machine translation only model probability distributions over possible target sentences of different lengths. Decoding the most probable sentence given the model is, however, a difficult task. Currently, we use heuristic search algorithms that try to compensate for the unpleasant fact that the number of theoretically possible target sentences grows exponentially with their length. If we had a reliable estimate of the target sentence length and could use it during the decoding, it might make it more efficient.

Co se naučíte: detaily fungování neuronového strojového překladu, práce s frameworky pro neuronový strojový překlad a hluboké učení

Skills you'll acquire: details of neural machine translation, working with frameworks for neural machine translation, and deep learning in general