DIUS

Veľké jazykové modely ako základ pre osobného digitálneho asistenta

Abstrakt

Spracovanie prirodzeného jazyka (NLP) bolo po celé desaťročia komplexným a nepolapiteľným cieľom v oblasti informatiky, často charakterizovaným pomalým pokrokom a čiastočnými riešeniami. Tento materiál popisuje vplyv veľkých jazykových modelov (LLM), umožnený pokrokom v oblasti hĺbkového učenia, dostupnosťou rozsiahlych jazykových dát a prístupným výpočtovým výkonom vo forme GPU. Analyzujeme integráciu LLM do každodenných aplikácií, so zameraním na ich význam pre slovenský jazyk a vývoj Jarvisa, nášho osobného asistenta využívajúceho tieto technológie.

Úvod

Spracovanie prirodzeného jazyka počítačmi je zložitá a netriviálna úloha, ktorá sa dlho považovala za nedosiahnuteľný cieľ vo výskume a priemysle, charakterizovaná pomalým pokrokom a čiastočnými, nedokonalými riešeniami. Táto oblasť sa etablovala ako špecializované vedecké pole známe ako počítačová lingvistika, NLP alebo všeobecnejšie jazykové technológie (LT). Aktuálne sme svedkami revolúcie v oblasti NLP spôsobenej proliferáciou veľkých jazykových modelov (LLM), ktoré ťažia z pokroku v oblasti hĺbkového učenia v kombinácii s dostupnosťou rozsiahlych jazykových dát a prístupným výpočtovým výkonom vo forme GPU.

Pozadie

Jazykové technológie sú už dávno prítomné v mnohých aspektoch nášho každodenného života, od prediktívneho vkladania textu na virtuálnych klávesniciach mobilných telefónov, kontroly pravopisu, a internetových vyhľadávačov až po prácu s rozsiahlymi jazykovými dátami alebo interakciu medzi človekom a počítačom. LLM prinášajú zásadnú zmenu, umožňujúc nielen komunikáciu s počítačmi v prirodzenom jazyku (písaním aj hovorením), ale v mnohých oblastiach aj prekonávajú schopnosti ľudí pri spracovaní jazyka a informácií.

Jazyková rôznorodosť a LLM

Najlepšie a najrozmanitejšie LLM existujú pre angličtinu, keďže väčšina výskumu a vývoja sa realizuje práve pre tento jazyk a existuje veľa voľne použiteľného anglického jazykového obsahu. Nasleduje skupina „veľkých európskych“ jazykov – nemčina, francúzština a španielčina, ako aj čínština a japončina. Väčšina ostatných (národných) európskych jazykov

je menej pokrytá, hoci hlavné komerčne dostupné jazykové modely dosahujú takmer bezchybnú gramatiku. LLM takto pomáhajú prekonávať "digitálnu medzeru" medzi "veľkými" jazykmi spracovanými na špičkovej úrovni a "malými" jazykmi (ako slovenčina), ktorých podpora výskumu v NLP bola dlhé roky na chvoste s veľkými medzerami.

Výzvy a príležitosti pre slovenčinu

Hoci slovenčina je podporovaná veľkými modelmi, jej podpora v oblasti otvorených a voľne dostupných (Open Source) modelov je veľmi nedostatočná a čiastková. Iba niekoľko týchto modelov obsahuje vôbec slovenčinu, aj to s nedokonalou gramatikou.

Asistent Jarvis

Na úspešné spracovanie slovenčiny vo vlajkových jazykových modeloch nadväzuje asistent Jarvis, ktorý má za cieľ slúžiť ako osobný asistent poskytujúci špecializovaný prístup k vybraným zdrojom a službám. Medzi jeho hlavné funkcie patrí asistovaná správa nákupov v predajni potravín Od našich, ovládanie inteligentnej domácnosti a do budúcna prístup k rôznym riešeniam poskytujúcim služby na báze umelej inteligencie. Komunikácia s asistentom Jarvis prebiha cez aplikáciu , ktorá umožňuje používateľom efektívnu komunikáciu. Vďaka integrácii s modernými technológiami a platformami môže Jarvis asistovať v širokej škále oblastí na základe preferencií používateľa.

Záver

Pokrok v oblasti LLM predstavuje významný míľnik v NLP, ponúkajúci bezprecedentné schopnosti a prístupnosť. Pre menšie jazyky ako slovenčina tieto modely nielen zlepšujú kvalitu spracovania jazyka, ale tiež otvárajú nové možnosti pre aplikácie, ktoré zlepšujú každodenný život. Asistent Jarvis je príkladom potenciálu týchto technológií, poskytujúc pohľad na budúcnosť, kde jazykové bariéry sú čoraz viac eliminované.

Referencie

Garabík, Radovan (2023). Language Report Slovak. In: Rehm, Georg, Way, Andy (eds) European Language Equality. Cognitive Technologies. Springer, Cham. https://doi.org/10.1007/978-3-031-28819-7_3

Šimková, Mária, Radovan Garabík, Katarína Gajdošová, Michal Laclavík, Slavomír Ondrejovič,

Jozef Juhár, Ján Genči, Karol Furdík, Helena Ivoríková, and Jozef Ivanecký (2012). Slovenský jazyk v digitálnom veku – The Slovak Language in the Digital Age. META-NET White Paper

Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/slovak