Majhni jezikovni modeli za brezkontekstno pridobivanje informacij v evropskih jezikih

---

Oznaka in naziv projekta

Z2-70067 Majhni jezikovni modeli za brezkontekstno pridobivanje informacij v evropskih jezikih
Z2-70067 Small Language Models for Zero-Shot Information Extraction in European Languages

Logotipi ARIS in drugih sofinancerjev

Projektna skupina

Vodja projekta: dr. Erik Novak

Sodelujoče raziskovalne organizacije: *Povezava na SICRIS*

Sestava projektne skupine: *Povezava na SICRIS*

Vsebinski opis projekta

Namen projekta je razvoj malih jezikovnih modelov (SLM) za namen ekstrakcije informacij brez nadzorovanega učenja v evropskih jezikih, s posebnim poudarkom na slovenščini. Projekt naslavlja tri ključne izzive obstoječih velikih jezikovnih modelov (LLM): visoke zahteve po računalniških virih, ki otežujejo lokalno namestitev pri manjših organizacijah; pomanjkljivosti pri učnih podatkih za občutljiva področja in jezike z malo viri; ter neskladnost izhodnih vrednosti, ki zmanjšuje zanesljivost ekstrakcije informacij. Za reševanje teh izzivov bomo razvili računsko nezahtevne modele, optimizirane za komercialno strojno opremo in lokalno namestitev, ustvarili obsežne večjezične referenčne podatkovne zbirke za področja z občutljivimi podatki, ter izvedli celovite evalvacije zmogljivosti modelov. Vsi rezultati projekta – modeli, podatkovne zbirke in programska koda – bodo javno dostopni (kadar bo to mogoče).

Faze projekta in opis njihove realizacije

Faza 1 – Zbiranje in priprava podatkov (WP1): Identifikacija ustreznih podatkovnih zbirk za evropske jezike (s poudarkom na slovanskih jezikih) iz virov, kot so Common Crawl in repozitorij Clarin.si; analiza kakovosti in jezikovne zastopanosti zbranih podatkov s pomočjo modelov za označevanje besednih vrst in identifikacijo jezikov; obogatitev in sintezno generiranje primerov za področja medicine in naravoslovja; ter formatiranje podatkovnih zbirk za pred-učenje in fino nastavitev modelov.

Faza 2 – Razvoj modelov (WP2): Analiza in razvoj arhitekture SLM, vključno z razvojem in integracijo naprednih komponent iz enkoder- in dekoder-arhitektur (po vzoru ModernBERT); analiza, primerjava in razvoj algoritmov tokenizacije (BPE, SentencePiece) za evropske jezike; ter integracija tokenizatorja in arhitekture modela za učenje v WP3.

Faza 3 – Učenje in evalvacija modelov (WP3): Učenje malih jezikovnih modelov z različnimi strategijami (maskiranje, napovedovanje naslednjega žetona) in uporaba tehnik kvantizacije (PTQ, QAT, GPTQ) za zmanjšanje velikosti modela; evalvacija na referenčnih podatkovnih zbirkah in primerjava z obstoječimi SLM in LLM na merah natančnosti, priklica in F1; ter javna objava najboljnih modelov, podatkovnih zbirk in programske kode na platformah HuggingFace in GitHub.

Bibliografske reference