Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Lukáš Chaloupský
Název práce	Automatic generation of medical reports from chest X-rays in Czech
Rok odevzdání	2022
Studijní program	Informatika **Studijní obor** Softwarové a datové inženýrství

Autor posudku	Rudolf Rosa **Role** vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

Student se věnoval tématu, které je převážně mimo jeho obor studia, oceňuji proto, že si dokázal nastudovat dostatek relevantní literatury a dostatečně se s tématem seznámit, takže předkládá práci, která kvalitně dosahuje vytyčeného cíle. Práci jsme často konzultovali a diskutovali možnosti postupu, avšak možné postupy obvykle navrhoval sám student.

Původním úmyslem bylo soustředit většinu úsilí na zdokonalování vlastního modelu pro generování popisků radiologických obrázků. V průběhu práce se ale ukázalo, že řešení některých dílčích problémů si vyžaduje výrazně větší úsilí, než jsme původně soudili, architektura výsledného modelu proto byla nakonec plně přejata z existující předchozí práce pro angličtinu a nebyla významně vylepšena. Přesto jsem přesvědčen, že výsledky sestaveného modelu jsou velmi dobré. Model sice v žádném případě není připraven na reálné nasazení v klinické praxi, ale vytyčený směr výzkumu považuji za velmi slibný pro budoucnost; vzhledem k tomu, že jde dle našich informací o první model, který tuto úlohu řeší pro češtinu, jde o důležitý první krok k budoucímu vývoji dokonalejších řešení této úlohy. Z toho zároveň i vyplývá obtížnost prvního prošlapávání cesty a objevování problémů, které je nutné vyřešit.

Zásadním problémem se ukázala nedostupnost vhodných dat pro češtinu. Velice oceňuji, že student věnoval obrovské úsilí komunikaci s různými zástupci českého medicínského sektoru, včetně zástupců několika nemocnic a několika představitelů české technologické startupové scény věnujících se medicíně. Přes počáteční slibné přísliby a přes mnohá zdlouhavá jednání se ale problém získání reálných českých dat v podobě popsaných radiologických snímků ukázal v rámci diplomové práce jako neřešitelný. Student však vhodným způsobem paralelně rozvíjel alternativní záložní řešení spočívající v automatickém strojovém překladu anglických dat, které nakonec bylo využito pro finální model. Zde přitom student musel překonat různé problémy, které vyvstaly při použití překladového systému na data z domény a ve formátu neodpovídající trénovacím datům modelu. Navázené kontakty s medicínskými odborníky se pak podařilo využít k důsledné evaluaci výsledného systému, kdy vygenerované popisky hodnotil zkušený radiolog.

Těžištěm práce se nakonec stala adaptace anglického jazykového modelu GPT-2 na češtinu a dále na české medicínské texty. Ta byla nutná, neboť model GPT-2 je klíčovou komponentou celkové architektury zodpovědnou za vlastní generování výstupních textů. Student zde vyšel z existujících postupů, nicméně pečlivě provedl mnoho experimentů zaměřených zejména na způsob sestavení a filtrování podkladových dat, tj. českých textů obecných a medicínských. Kvalita výsledného modelu je dobrá, jde tedy o jakýsi původně nezamýšlený vedlejší výstup práce, který ale z pohledu praktické využitelnosti může být užitečnějším než vlastní výsledek celé práce – tyto velké jazykové modely jsou užitečné v mnohých aplikacích, přičemž jejich dostupnost pro jiné jazyky než je angličtina je značně omezená.

Navrhuji práci hodnotit známkou výborně.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 24. 8. 2022

Podpis: