

Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce František Trebuňa
Název práce Generating text from structured data
Rok odevzdání 2021
Studijní program Informatika
Studijní obor General Computer Science

Autor posudku Mgr. Rudolf Rosa, Ph.D. Vedoucí
Pracoviště Ústav formální a aplikované lingvistiky

K celé práci

lepší OK horší nevyhovuje

	lepší	OK	horší	nevyhovuje
Obtížnost zadání	X			
Splnění zadání	X			
Rozsah práce <i>... textová i implementační část, zohlednění náročnosti</i>	X			
<p>Autor se ve své práci věnuje obtížnému problému, který zatím v existující literatuře není uspokojivě vyřešen, jde tedy o ambiciózní úlohu.</p> <p>Generování textového popisu sportovního utkání je běžným podtypem obecné úlohy generování textu ze strukturovaných dat, a užitý dataset RotoWire je pro tuto úlohu de facto standardem. Autor důkladně prostudoval řadu existujících souvisejících vědeckých prací, přičemž klíčové předchozí práce v textu popisuje a vysvětluje; celkem autor cituje asi 30 odborných prací, což je pro bakalářskou práci nadprůměrné množství.</p> <p>Autor následně navrhuje, implementuje a porovnává několik metod řešení úlohy včetně některých variací těchto metod, přičemž vychází z existujících prací, ale navrhuje i vlastní úpravy, se kterými experimentuje. Z tohoto důvodu také autor obvykle nepřejímá zdrojové kódy předchozích prací, ale reimplementuje je sám, což je pracnější, ale umožňuje mu to snáze tyto metody modifikovat. Autor přitom vychází z moderních postupů založených na rekurentních umělých neuronových sítích s řadou pokročilejších komponent (attention, beam search, copy mechanism, apod.), což jednoznačně daleko přesahuje rámec bakalářského (a částečně i magisterského) studia na MFF.</p> <p>Značnou pozornost věnuje autor zpracování a reprezentaci datasetu a zejména vstupních tabulkových dat. Jak autor v práci zmiňuje, dataset RotoWire má mnoho problémů, které si autor uvědomuje a částečně je řeší; přesto se jedná o standardní dataset pro tuto úlohu, což je důvodem jeho použití v této práci. Autor nejprve aplikuje některé běžné metody předzpracování dat s vlastními úpravami (např. Byte Pair Encoding, desambiguace a unifikace pojmenovaných entit). Následně autor experimentuje s několika způsoby uspořádání a filtrování vstupních tabulek; jde svým způsobem o pravidlové či heuristické varianty plánování obsahu (content planning), které autor v práci využívá i v jeho variantě založené na strojovém učení.</p> <p><i>(pokračování na další straně)</i></p>				

Autor experimenty důsledně vyhodnocuje, a to jak automaticky, tak manuálně. Hlavní přínos práce dle mého názoru spočívá v porovnání jednotlivých metod a jejich variant, kdy autor ukazuje, které metody přinášejí jaké zlešení výsledků, a to jak kvantitativně, tak kvalitativně. Autor si při manuálním vyhodnocení zejména všímá vlivu jednotlivých metod na repetitivnost a halucinaci, což jsou dva nejběžnější a nejzásadnější problémy neuronových přístupů k dané úloze. Autor uzavírá, že ačkoliv se mu podařilo dosáhnout podstatně lepších výsledků než základní metoda (baseline), výsledné texty stále trpí mnoha problémy, a zamýšlí se nad nejslibnějšími postupy pro dosažení dalších zlepšení. Přesto autorem představené výsledky považuji za dobré a jím dosažené zlepšení za podstatné.

Celkově práce jednoznačně splňuje nároky kladené na bakalářskou práci a zejména svým výzkumným a experimentálním rozsahem se blíží spíše práci magisterské.

Textová část práce

lepší OK horší nevyhovuje

Formální úprava <i>... jazyková úroveň, typografická úroveň, citace</i>	X			
Struktura textu <i>... kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu</i>	X	X		
Analýza	X	X		
Vývojová dokumentace		X		
Uživatelská dokumentace		X		

Práce je psaná dobrou angličtinou, obsahuje všechny důležité části, správně cituje zdroje, má velmi dobrou typografickou úpravu. Použité postupy jsou kvalitně vysvětleny pomocí textových popisů, vzorců a diagramů. Práce na vhodných místech prezentuje i ukázky dat a výstupů, což dobře ilustruje řešený problém a dosažené výsledky.

Vývojová a uživatelská dokumentace je stručná, ale shrnuje nejdůležitější informace, což je pro práci výzumného typu zcela postačující.

Implementační část práce

lepší OK horší nevyhovuje

Kvalita návrhu <i>... architektura, struktury a algoritmy, použité technologie</i>	X			
Kvalita zpracování <i>... jmenné konvence, formátování, komentáře, testování</i>	X	X		
Stabilita implementace	X	X		

Autor používá vhodné nástroje, postupy popsané v práci dobře implementuje, vychází z programátorských konvencí běžných pro práci s umělými neuronovými sítěmi.

Autor k práci přikládá sadu zdrojových kódů v Pythonu, strukturu práce vysvětluje v příloze a v souboru README, vlastní skripty obsahují dokumentaci. Zároveň řešení obsahuje několik řeckněme uživatelských skriptů, které dle návodu v příloze práce umožňují odbornému uživateli spuštění jednotlivých metod.

Úroveň implementační části považuji za obvyklou až mírně nadprůměrnou pro tento typ prací, t.j. pro práci orientovanou výzkumně.

Celkové hodnocení Výborně
Práci navrhuji na zvláštní ocenění Ne

Datum

Podpis