Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Kristína Szabová		
Název práce	Neural Concept-to-text Generation with Knowledge Graphs		
Rok odevzdání	2023		
Studijní program	Informatika	**Studijní obor**	Jazykové technologie a počítačová lingvistika

Autor posudku	Jindřich Libovický	**Role**	oponent
Pracoviště	ÚFAL MFF UK		

Text posudku:

Předložená práce se věnuje problému, jak zohlednit všední zkušenost a znalost (commonsense knowledge) při generování jazyka pomocí velkých jazykových modelů. Velké jazykové modely v posledních měsících a letech zaznamenaly značný pokrok, stále ale mají problémy s usuzováním o všedních záležitostech (commonsense reasoning), což často vede chybám při generování textu. Schopnost takového usuzování v této oblasti se měří (mimo jiné) pomocí datasetu CommonGen, na kterém autorka obhajované práce demonstruje efektivitu ověřovaných experimentálních metod.

Práce má celkem 87 stran, hlavní text začíná na straně 2 a končí na straně 53, z toho deset stran obsahuje výhradně tabulky s příklady generovaného textu. Práce se skládá z 6 kapitol včetně úvodu a závěru. Kapitola 1 (7 stran) vysvětluje základní koncepty pro pochopení řešení úlohy. Kapitola 2 (11 stran) jde o krok zpět a vysvětluje základní koncepty pro zpracování přirozeného jazyka pomocí neuronových sítí. Kapitola 3 (6 stran) popisuje provedené experimenty a kapitola 4 (4 strany) představuje výsledky experimentů.

Text práce je velmi dobře strukturovaný. Práce je psaná dobrou angličtinou bez zjevných gramatických chyb, ale často se zvláštním výběrem slov nebo nepřesnými formulacemi (např. nesprávné ,,perform experiment“ místo ,,conduct experiment“; ,,explore improvements“ neříká, jestli se jedná o už existují zlepšení nebo zlepšení navrhovaná autorkou; ,,advantageous“ použité ve smyslu ,,useful“; ,,method has two components“ namísto ,,two steps“).

Kapitola 1 dobře vysvětluje, jaké úloze se autorka ve své práci věnuje a proč má smysl takovou úlohu řešit.

Kapitola 2 je nejslabším místem práce. Začíná tím, že vysvětluje historické postupy, které nejsou pro práci relevantní. Následuje popis architektury Transformer, který je poněkud zmatený. Autorka se snaží základní myšlenku architektu vysvětlit pomocí rozdílu od attention mechanismu v rekurentních encoder-decoder modelech. To se ale týká pouze tzv. cross-attentionu mezi en-

kodérem a dekodérem, jehož role zůstává stejná. Zásadní role self-attentionu není poskytování rozhraní mezi vstupem a výstupem, ale to že je odpovědný za reprezentaci kontextu v modelu. Nahrazuje tak rekurenci v předchozích modelech, se kterými se autorka pokouší architekturu porovnat. Na straně 17 náhle začíná popis dekódování (což je stěžejní koncept pro generování textu pomocí jazykových modelů), aniž by se předtím vysvětlilo, co to dekódování je: tedy, že modely počítají pravděpodobnost tokenů a potřebujeme algoritmus, který z těchto pravděpodobností udělá textový výstup. Vzápětí následuje formální matematický popis mechanismu self-attention, který je není zcela dobře. Za vstup považuje vektor reálných čísel – vstupem ale musí být matice, jinak by modely reprezentovaly vstupní tokeny izolovaně bez kontextu. Další části (popis knowledge bases a Graph Neural Networks) jsou lepší, ale stále působí poněkud nejistě.

Kapitola 3 popisuje původní experimentální práci autorky. Experimenty jsou popsány dobře a text netrpí podobnými nedostatky jako v předchozí kapitole. Z textu je zřejmé, jak se jednotlivé experimenty liší a jak na sebe navazují.

Kapitola 4 prezentuje výsledky experimentů, nejprve kvantitativně pomocí automatických metrik, následně kvalitativně na základě manuálního hodnocení. Práce bohužel nevyužívá dostatečně potenciál automatických metrik. Pro často používané metriky jako je BLEU existují nástroje, které pomáhají manuální evaluaci, například tím, že dovedou vyhledávat n-gramy, které vedly k lepšímu skóre. S jejich pomocí je pak možné zpětně určit, jestli vyšší skóre v automatické metrice skutečně znamená i vyšší kvalitu výstupu.

Část o manuální evaluaci nezahrnuje důležité informace o procesu hodnocení. Není zmíněné, kdo byli v tomto případě hodnotící. Pokud tím, kdo výstupy systému hodnotil/a, byla autorka sama, bylo by nutné zabránit tomu, aby hodnocení bylo ovlivněné apriorními představami o tom, jak jednotlivé metody fungují, například tím, že by se vygenerované věty hodnotily v náhodném pořadí.

Závěr shrnuje hlavní poznatky z experimentální části práce, tj. že přidání informací z knowledge base zlepšuje kvalitu generování textu oproti dotrénování předtrénovaného jazykového modelu.

Autorka předloženou prací prokázala, že se orientuje v problematice generování textu pomocí jazykových modelů. Dále práce ukazuje, že autorka dovede plánovat, implementovat a vyhodnotovat výpočetní experimenty s jazykovými modely.

- *Silnými stránkami práce jsou:* dobře motivované, dobře navržené a dobře provedené experimenty, jejichž výsledky dobře ukazují slabé stránky současných jazykových modelů a další směr, kterým je možné se při řešení podobných úloh ubírat.

- *Slabými stránkami práce jsou:* relativně kratší rozsah, místy horší kvalita textu a metodologické nedostatky v manuálním hodnocení.

Otázky k práci

- Myslíte si, že problémy s manuálním hodnocením, které zmiňuje tento posudek, negativním způsobem ovlivnily věrohodnost manuálního hodnocení?
- Myslíte si, že metody, se kterými experimentujete v této práci jsou přenositelné do jiných oblastí generování přirozeného jazyka (např. asistenti typu ChatGPT)?

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 14. 8. 2023

Podpis: