Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Kristýna Neumannová		
Název práce	German Compounds in Transformer Models		
Rok odevzdání	2023		
Studijní program	Computer Science	**Studijní obor**	Language Technologies and Computational Linguistics

Autor posudku	Daniel Zeman	**Role**	Oponent
Pracoviště	Ústav formální a aplikované lingvistiky		

Text posudku:

Předložená diplomová práce zkoumá produkci složených podstatných jmen při strojovém překladu z angličtiny do němčiny. Cílem je zjistit, zda jsou strojové překladače schopné generovat složená slova ve srovnatelné míře jako lidé, zejména zda dokážou generovat složeniny, které se nevyskytly v trénovacích datech. Toto téma je zajímavé, protože pomáhá osvětlit aspekty chování algoritmů strojového překladu, které nezachytí běžně používané metriky jako BLEU, nicméně jsou pro němčinu a podobné jazyky důležité.

Práce se skládá z pěti kapitol. První kapitola přináší přehled dříve publikovaných prací a nástrojů jak pro samotný strojový překlad pomocí neuronových „transformerů“, tak pro neřízené dělení slov na menší jednotky a pro generování složených slov z takových jednotek. Druhá kapitola popisuje přímo nástroje a data použitá v předkládané diplomové práci. Třetí kapitola zkoumá výskyt složených substantiv ve výstupech systémů, které se účastnily každoroční soutěže ve strojovém překladu WMT v roce 2021. Čtvrtá kapitola stručně popisuje překladové modely natrénované přímo pro účely této práce, pátá kapitola pak zkoumá výskyt složených substantiv ve výstupech těchto modelů. Strukturu textu hodnotím jako dobře zvolenou. Práce má 60 stran, z nichž 26 je věnováno popisu autorčiny vlastní práce v kapitolách 3 až 5. Práce je psána slušnou a dobře srozumitelnou angličtinou s minimem překlepů.

Text práce slibuje, že zdrojový kód i výstupy analýzy složených slov jsou dostupné na Gitlabu, na uvedené adrese ale v době posuzování práce žádný obsah dostupný nebyl.

Konkrétní připomínky a otázky:

Strana 17: Na tomto místě by bylo užitečné mít nějaké příklady platných německých složených slov, která GermaNet nezná (byť se nějaké takové příklady objeví později v kapitole 3.3.1 mezi slovy vygenerovanými jedním z překladačů a neznámými z trénovacích dat). Lze nějak odhadnout, jaké je pokrytí GermaNetu, tj. jak často se v německém textu vyskytne složené slovo, které GermaNet nezná?

Strana 18: Poslednímu odstavci („Although we sorted the table according to…“) vůbec nerozumím. Chápu, že stejný koncept může být přeložen pomocí různých složených slov, z nichž jedno se objeví v referenčním překladu a druhé ne. Jak s tím ale souvisí počet vět, ve kterých se vyskytlo složené slovo? To je přece úplně jiná míra!

Strana 21: „The dots … represent the average number of systems that produced compounds for the frequency interval…“ Co to znamená? Aby byl systém započítán do dotyčného počtu, musel vygenerovat některé složeniny z daného frekvenčního intervalu? Nebo všechny? A všechny instance, nebo všechny typy?

Strana 25: „complex compounds that seemed to have no sense“ … Co přesně znamená, že seemed to have no sense? Jak *Sanktionsüberwachungsteam*, tak *Passagierlokalisierungsformular* jsou korektně utvořená složená slova s jasným významem.

Obecně: Práce zjevně vychází z předpokladu, že v případě složených slov více znamená lépe. Je to ale pravda? Kromě v práci zmíněné korelace (lepší systémy jsou současně ty, které produkují více složených slov), je k dispozici nějaké vyhodnocení, kde se bude přímo měřit přesnost, úplnost a F-skóre složenin vůči referenčnímu překladu?

Jak často strojový překlad vyrobí složeninu, kterou nevyrobí člověk? A je taková složenina alespoň někdy správně? (Tato otázka se týká jak systémů z WMT21 v kapitole 3, tak vlastních systémů autorky v kapitolách 4 a 5.)

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 29. května 2023 **Podpis**