Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Dominik Macháček
Název práce	Obohacování neuronového strojového překladu technikou sdíleného trénování na více úlohách
Rok odevzdání	2018
Studijní program	Informatika **Studijní obor** Matematická lingvistika

Autor posudku	RNDr. Ondřej Bojar, Ph.D. **Role** vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

Diplomová práce Dominika Macháčka studuje možnosti obohacení vstupu neuronových modelů strojového překladu s cílem využít lingvisticky relevantní (automatické) anotace a zlepšit kvalitu překladu.

Dominik pracuje s jedním z nejnovějších a v současné době i nejnadějnějších modelů, tzv. Transformerem, a zkoumá úspěšnost tzv. multi-task trainingu, tj. trénování modelu na více úlohách současně. Jen jedna z úloh (překlad) je hlavní, ale znalost dalších úloh by do modelu měla vnést potřebnou generalizaci a robustnost. V souladu se zadáním se Dominik soustředí především na nejjednodušší možnost zapojení dodatečné informace: dodatečné úlohy (např. morfologické značkování) jsou reprezentovány stejnou formou jako vstup a výstup pro překlad, tj. jako řetězec tokenů. Trénovací data pak proloženě obsahují obě úlohy.

Druhá zkoumaná metoda implementuje koncovou část neuronové sítě (dekodér) vícekrát, odděleně pro každou z úloh. Trénování probíhá na obou úlohách současně a základní úloha postupuje bez zpomalení (dle počtu trénovacích kroků). Tato metoda přirozeně vyžadovala větší množství implementační práce.

Kladně hodnotím zejména šíři provedených experimentů a jejich velmi pečlivou diskusi. Dodatečné úlohy na jednu stranu mohou pomoci, na druhou stranu nutně stojí určitý trénovací čas. Srovnáním s falešnými úlohami (počítání délky věty ap.) Dominik dokládá, že lingvistická informace pro překlad relevantní je, ale bohužel její užitečnost nevyváží dodatečné náklady na trénování. Podobně pečlivou evaluaci nacházíme i ve druhém přístupu, kde Dominik přechází k jiné základní implementaci modelu, OpenNMT místo Tensor2tensor. Kromě samotných experimentů s více dekodéry Dominik usiluje i o konfiguraci základního běhu OpenNMT, která bude co nejporovnatelnější s během Tensor2tensor.

Výsledky provedených experimentů jsou převážně negativní, s výjimkou pokusů na poměrně

malých datech (do 500 tisíc paralelních vět). Dominikova práce je tak hodnotná, protože naznačuje, že podobné dosud publikované práce jsou velmi pravděpodobně užitečné jen ve velmi omezených podmínkách a nikoli v případě rozsáhlejších trénovacích dat.

Práce je psána dobrou angličtinou s relativně malým počtem gramatických chyb. Struktura je přehledná. Text je dobře srozumitelný a je podrobně ilustrován, aby bylo jasné, jakou informaci model nově dostává.

Za zmínku stojí skutečnost, že přípravné experimenty, které ani nejsou v diplomové práci uvedeny, odhalují zajímavou citlivost modelu na explicitní zvýraznění konců vět a podařilo se je publikovat v článku na TSD 2018.

Jako vedoucí jsem s Dominkovou prací i jejím výsledkem spokojen a předloženou diplomovou práci jednoznačně doporučuji k obhajobě.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 10. 9. 2018

Podpis: