POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Intent Detection Module for a Conversational Assistant**
Jméno autora:	**Daria Ozerova**
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra kybernetiky
Oponent práce:	Ing. Jan Pichl
Pracoviště oponenta práce:	Katedra kybernetiky

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce.

Průměrně náročné zadání týkající se rozpoznávání intentu, které však kromě analýzy samotných algoritmů obsahuje i analýzu potřebných zdrojů a nasazení modelu do produkčního řešení.

Splnění zadání **splněno**

Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Zadání splněno.

Zvolený postup řešení **správný**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Zvolený postup řešení zahrnuje použití transformer-based modelů, které dosahují state-of-the-art výsledků na širokém spektru NLP úloh. V práci je také uvedena analýza výhod a nevýhod některých konkrétních architektur i s ohledem na reálný provoz těchto modelů.

Odborná úroveň **A - výborně**

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

Práce je na vysoké odborné úrovni a detailně popisuje přístupy k řešení problému rozpoznávání intentu v dialogových systémech. Popsány jsou supervised i unsupervised metody včetně hlavních výhod a nevýhod. V sekci 2.2.4 bych v porovnání embeddingů uvítal i word2vec, když už je zmíněn v předchozí části jakožto typický zástupe reprezentace slov.

Formální a jazyková úroveň, rozsah práce **A - výborně**

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

Formální a jazyková stránka práce je v pořádku. Pouze bych vytkl dvě drobnosti. První je často vídaná chyba v LaTeXu nesprávného používání znaku pro jednotku palce (") místo znaku pro uvozovky (“). Druhou věcí je používání zájmena *já* místo v odborných pracích běžněji používaného *my* (i přesto, že se jedná o individuální práci).

Výběr zdrojů, korektnost citací **A - výborně**

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

V práci jsou korektně citovány relevantní zdroje týkající se úlohy rozpoznávání intentu i zmíněných architektur modelů.

Další komentáře a hodnocení

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.
Vložte komentář (nepovinné hodnocení).

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Práce se zabývá problematikou rozpoznávání intentu v konverzačních systémech, což je jedna z nejzákladnějších částí těchto systémů. Jednotlivé porovnávané přístupy a architekty modelů patří sice mezi běžně používané pro tyto typy úloh, jsou ale navíc v této práci podrobně rozebrány nejen z hlediska přesnosti samotných modelů, ale i z hlediska použití v reálném systému. Na práci oceňuji, že se problematikou zbývá komplexně a řeší i často opomíjené věci jako je časová a paměťová náročnost inference modelů, updatování natrénovaných modelů o nová data, celkový životní cyklus modelu i nasazení pomocí inferenčního serveru.

Otázky k obhajobě:

1. Jaká je velikost batche v inferenci na CPU v tabulce v sekci 2.2.5
2. Plánujete porovnávat výkon modelů běžících v jiných inferenčních serverech (TFServing, Nvidia Triton)?

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **A - výborně.**

Datum: 1.6.2022 Podpis: