# Posudek diplomové práce

**Matematicko-fyzikální fakulta Univerzity Karlovy**

| | |
|---|---|
| **Autor práce** | Bc. Petra Vysušilová |
| **Název práce** | Czech NLP with Contextualized Embeddings |
| **Rok odevzdání** | 2021 |
| **Studijní program** | Informatika **Studijní obor** Umělá inteligence |

| | | | |
|---|---|---|---|
| **Autor posudku** | Milan Straka | **Role** | Vedoucí |
| **Pracoviště** | Ústav formální a aplikované lingvistiky | | |

**Text posudku:**

Hluboké učení dosáhlo v uplynulé dekádě znatelných úspěchů v mnoha oblastech. Také při zpracování přirozeného jazyka jsou v posledních přibližně pěti letech nejlepší metody postavené na hlubokých neuronových sítích. Nedávný největší pokrok je použití tzv. kontextualizovaných embeddingů, které dovolují efektivní přenos znalostí z modelů předtrénovaných modelováním čistého textu. Cílem diplomové práce bylo prozkoumat různé varianty použití kontextualizovaných embeddingů při automatickém zpracování češtiny.

První kapitola diplomové práce popisuje metody hlubokého učení a vývoj vedoucí k návrhu kontextualizovaných embeddingů. Dále pak popisuje model BERT, což je jedna z nejúspěšnějších variant těchto embeddingů, a osm nejznámějších úprav tohoto modelu. Tato kapitola je první přínos diplomové práce, protože vzniklých 29 stran textu je velmi kvalitních a pěkným způsobem shrnuje vývoj a poznatky v této překotně se rozvíjející oblasti až do současnosti.

Zbylá část diplomové práce se věnuje provedeným experimentům ve třech oblastech – morfologického značkování (určení slovních druhů a podrobných morfologických kategorií), lemmatizaci a analýzy sentimentu (pozitivity / negativity daného dokumentu). V případě morfologické analýzy a lemmatizace řešitelka navázala na existující prototyp, který přepsala pro použití s TensorFlow 2 a dramaticky rozšířila jeho možnosti. Poté provedla přes 40 experimentů s různými variantami modelu BERT, způsoby použití kontextualizovaných embeddingů a různými hyperparametry, a překonala nejlepší dosažené výsledky v těchto úlohách. Zároveň jsou výsledky zajímavé i z vědeckého hlediska – řešitelka na rozsáhlém porovnání jako (dle mých nejlepších znalostí) první ukazuje, že v diskutované úloze je nejvhodnější způsob použití kontextualizovaných embeddingů jiný než běžně používaný. Ve třetí zmiňované úloze, analýze sentimentu, provedla řešitelka opět rozsáhlé vyhodnocení hyperparametrů a dosáhla nejlepších výsledků na dvou ze třech použitích datasetů.

Nejlepší natrénované modely a použitá implementace jsou veřejně k dispozici.

Práci považuji za zdařilou, věnující se aktivně zkoumanému tématu, prokazující jak porozumění aktuálním metodám hlubokých neuronových sítí tak zároveň technickou způsobilost k jejich využití, včetně trénování na GPU akcelerátorech.

**Práci** doporučuji **k obhajobě.**

**Práci** nenavrhuji **na zvláštní ocenění.**

| |
|---|
| |

**Datum** 23. srpen 2021 **Podpis**