

# Posudek oponenta bakalářské práce

Autor/autorka práce: **Martin Matas**

Název práce: **Rozpoznávání pojmenovaných entit pomocí neuronových sítí**

Obsah práce

V práci je provedeno srovnání několika existujících nástrojů pro rozpoznávání pojmenovaných entit. Na základě výsledků porovnání a požadovaných vlastností je vybrána implementace LM-LSTM-CRF. Hlavním důvodem je možnost přidání nových příznaků, což je největším přínosem práce. Byly přidány 3 typy příznaků a byla otestována úspěšnost systému s jejich různými kombinacemi. Testování bylo provedeno pro čtyři jazyky na standardních datových sadách.

Kvalita řešení a dosažených výsledků

Práce je hlavně rozšířením stávajícího systému, který je volně dostupný na službě GitHub. Rozsah vlastní implementace je menší a zahrnuje zejména přidání nových příznaků a pak také další úpravy existujícího kódu. Vytvořený kód je přehledný a dostatečně komentovaný. Software je možné bez problémů spustit a otestovat jak na osobním počítači tak na Metacentru.

Výsledky práce

Přidání příznaků přineslo zlepšení oproti původnímu systému. Bylo dosaženo výsledků, které jsou srovnatelné s dostupnou literaturou. Nicméně systém, který byl vybrán, dosahuje nejnižších úspěšností ze čtyř testovaných systémů a byl vybrán hlavně kvůli snadné rozšiřitelnosti. Je tedy otázka zda a o kolik by ostatní systémy fungovaly lépe.

Formální úroveň

Práce je přehledně strukturována a téměř neobsahuje překlepy nebo pravopisné chyby. V textu jsou následující nedostatky:

- str. 9 - V chybové matici jsou prohozeny FP a FN, v textu jsou zmíněny dva druhy chyb (TN, FN), TN ale není chyba.
- U popisu korpusů není uvedeno, jak jsou rozděleny - počty dat pro trénování, validaci a testování
- použití desetinné tečky místo čárky
- Ve výsledcích jsou uváděny hodnoty přesnosti, úplnosti a F-míry, F-míra ale neodpovídá hodnotám přesnosti a úplnosti, ze kterých je počítána. Místo přesnosti je zřejmě uváděna „accuracy“ vrácená systémem.
- Číslování stran není od úvodu.

Práce využívá velké množství odborné literatury. Použité prameny jsou dostatečně citovány.

Splnění zadání

Zadání bylo splněno bez výhrad.

# Posudek oponenta bakalářské práce

Dotazy k práci

Vysvětlíte, proč má pořadí příznaků vliv na výsledky klasifikace (str. 40, tabulka 5.7).

Vysvětlíte výpočet F-míry a proč se ve výsledcích liší od harmonického průměru přesnosti a úplnosti.

Zdůvodněte obtížnost přidání nových příznaků do ostatních systémů.

Jak byly počítány výsledky, jedná se vždy o jeden běh nebo je konečný výsledek průměrován z několika běhů programu?

Jaké nástroje byly použity pro stemming a lemmatizaci?

Navrhuji hodnocení známkou **velmi dobře** a práci doporučuji k obhajobě.

V Plzni 16.5.2018

Ing. Ladislav Lenc, Ph.D.

