

Posudek oponenta diplomové práce

Bc. Josef Strolený
Vícejazyčné metody pro hledání sémantické podobnosti slov

Diplomant se ve své práci zabývá metodami pro určování vícejazyčné podobnosti slov. Jedná se o velice aktuální téma a je nutno zdůraznit, že pro diplomanta se jedná o velice náročnou problematiku. V teoretické části práce jsou rozebírány nejprve jednojazyčné metody pro určování podobnosti slov a poté jejich vícejazyčné modifikace.

Práce má dobrou logickou strukturu, text se příjemně čte, a přesto je dostatečně věcný. Práce je doplněna velkým množstvím užitečných názorných obrázků. K obsahu práce mám jen několik výhrad. Velká část práce je věnována metodám založeným na sémantických sítích a tyto metody nejsou vůbec v praktické části použity. Bag-of-words hypotéza je definovaná nesprávně, navíc u ní chybí reference.

V práci mi chybí jasné vymezení cíle. V závěru se pak student neměl o co opřít a nezdůraznil dostatečně přínosy práce. Po přečtení úvodu a závěru práce budí dojem, že jejím jediným cílem a poté i údajným přínosem je vylepšení stávajících systémů pro určování vícejazyčné sémantické podobnosti slov jejich pouhou kombinací. Tento cíl dle mého názoru splněn nebyl (viz zhodnocení kvality dosažených výsledků). Práce má však řadu jiných přínosů, které jen nejsou v textu dostatečně zdůrazněné. V popisu implementovaných algoritmů mohl jít autor často do větší hloubky. U většiny metod je shrnut jen základní princip bez větších detailů.

I přes výše zmíněné prohřešky je po obsahové stránce práce velice zdařilá.

Formální úroveň práce je velmi dobrá, ale text práce obsahuje ne úplně zanedbatelné množství překlepů, občasné špatné skloňování, kde autor zapomene, co je podmětem a dokonce několik hrubých chyb. Vzhledem k rozsahu práce lze ale tyto případy označit za ojedinělé. Formátování odstavců je nevhodné, je nastavena jak meziodstavcová mezera, tak odsazování prvního řádku. Obzvlášť u jednořádkových odstavců to působí zvláštně.

Práce obsahuje dostatečné množství relevantních zdrojů, a to jak tištěných, tak webových. Jako drobný nedostatek vidím, že v některých případech není citován autor algoritmu, ale nějaký pozdější článek, který daný algoritmus „převyprávěl“. Příkladem je „Wu & Palmerův algoritmus“. Také formát citací je nesprávný. Autor si evidentně psal bibtex ručně a pro vědecké články zvolil špatný druh. Důsledkem toho jsou v citacích konferenčních článků uvedeni pouze autoři, rok a název článku.

Práce prezentuje zlepšení o 6% oproti state-of-the-art. Uvedené „state-of-the-art“ výsledky jsou ale o více než 12% horší než výsledky publikované na soutěži SemEval 2017. Dále diplomant optimalizuje hyperparametry na testovacích datech, čímž dostává neférové srovnání. Nicméně v rámci práce bylo otestováno poměrně velké množství metod. Byla vytvořena slušně fungující metoda s velkým slovníkem pro jazyky, kterými se zabývá naše výzkumná skupina. Výsledky práce tedy budou velice pravděpodobně dále využity. Práci tedy hodnotím jako velice přínosnou, ačkoliv její přínosy nejsou dostatečně důrazně prezentovány v textu práce. Příložený program je velice těžké

spustit, i když to z větší části není diplomantova chyba, protože některá potřebná data není možné automaticky stáhnout ani nahrát na médium.

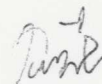
Zadání bylo splněno bez výhrad.

Práci doporučuji k obhajobě a hodnotím klasifikačním stupněm

„výborně“.

Doplující otázky:

1. Pokoušel jste se zjistit, proč jsou vámi naměřené výsledky state-of-the-art systému o 12% horší, než výsledky publikované na SemEval 2017, když jde podle vašich slov o tentýž systém a používáte shodná testovací data?



Ing. Ondřej Pražák
(oponent DP)

V Plzni 4. června 2018

**SOUHLASÍ
S ORIGINÁLEM**



Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky

①