Posudek oponenta bakalářské práce

Autor práce: Jiří Hankovec

Název práce: Testování podobnosti vět

Obsah práce:
Práce se zabývá automatickým určováním podobnosti dvou vět z hlediska významu (z angl. semantic textual similarity - STS). Vstupem této úlohy jsou dvě věty a výstupem je jejich podobnost. Cílem práce bylo vytvoření českého korpusu pro testování STS algoritmů a dále adaptace UWB systému [1] na český jazyk. UWB systém byl vytvořen NLP skupinou na katedře KIV za účelem účasti na soutěži SemEval 2016, konkrétně na úloze č. 1. Práce je výzkumného charakteru a od studenta vyžadovala proniknout do teoreticky poměrně náročné oblasti zpracování přirozeného jazyka.

Kvalita řešení a dosažené výsledky:
Student použil část anglických korpusů poskytovaných soutěží SemEval z roku 2014 a 2015 a páry vět přeložil do češtiny (celkem 1200 párů vět). Stávající UWB systém rozšířil o předzpracování textů vhodné pro češtinu a o normalizaci slovních tvarů (důležité pro jazyky s bohatou morfologií). Na přeloženém korpusu českých vět tento systém otestoval a dosáhl Pearsonovy korelace přibližně 0,8, což je uspokojivý výsledek.

Formální úroveň práce:
Bakalářská práce se skládá z 37 stran vlastního textu (49 stran včetně úvodních stránek, referencí a přílohy). Práce je vysázena v LaTeXu. Z práce je na první pohled zřejmé, že byla dokončována na poslední chvíli. Struktura dokumentu a popis jednotlivých částí je poměrně chaotický a často nepřesný. Některé pasáže jsou velmi vágní a zbytečné. U mnohých tvrzení chybí citace. Student věnuje většinu prostoru popisu stávajícího UWB systému a velice málo prostoru práci vlastní. Hlavní výtka je, že práce neobsahuje kapitolu o vytvořeném korpusu. Na pár místech v práci je korpus zmíněn, ale jelikož se jedná o hlavní přínos práce, mělo by popisu korpusu a jeho vytvoření být věnováno podstatně více prostoru. Další nepřesnosti:

- Nekonzistence ve vzorcích - notace (velká a malá písmena), násobení jednou jako AB a podruhé A x B, atd. Vzorec 3.6 je špatně.
- Celá kapitola 3.1 Strojové učení je příliš vágní a zbytečná – student by měl popisovat strojové učení ve vztahu k STS, ne obecně.
- Předzpracování dat je popsané v metodách STS - mělo by být jinde.
- Kapitola 3.3 Metody založené na znalostních bázích je zbytečná - není řečena jediná metoda pro STS.
- Není definováno tf-idf pro znaky – čtenář si pravděpodobně domyslí, ale mělo by to tu být.
- U LSA by měla být citace [2]. Chybí základní myšlenka LSA - redukce dimenze.
- Jsou používány 3 metody regrese, ale vysvětlena jenom lineární regrese.
- Vzorce 4.1 a 4.4 jsou stejné v případě ze se střední hodnoty a rozptyly počítají na vzorku dat (nejsou dopředu známy) – což je případ této práce.
- V kapitole 5 se opakují věci již popsané v kapitole 3.
- Tabulka 6.1 neukazuje kombinace metod (to ukazuje tabulka 6.2), ale jednotlivé metody.
- Obr. 6.1 – jaký systém je tady použit? S jakým krokem byl tento graf počítán? Při dostatečně malém kroku rozhodně graf nebude monotónní.
- V tabulce 6.3 jsou porovnávány výsledky mezi EN a CS na úplně jiných datech – čili jsou neporovnatelné. Proč nejsou porovnané na stejných datech, když se jednalo o překlad?
-

Splnění zadání:
Práce splňuje zadání. Navrhuji hodnocení známkou **velmi dobře** a práci doporučuji k obhajobě.

V Plzni 18. 8. 2017

Ing. Tomáš Brychcín, Ph.D.

Reference
[1] T. Brychcin and L. Svoboda. UWB at SemEval-2016 Task 1: Semantic Textual Similarity using Lexical, Syntactic, and Semantic Information. SemEval@NAACL-HLT, 2016.
[2] T. Landauer and P.W. Foltz and D. Laham. Introduction to Latent Semantic Analysis. Discourse Processes, 1998.