CHARLES UNIVERSITY PRAGUE

faculty of mathematics and physics

STU – Slovenská Technická Univerzita

FIIT - Fakulta Informatiky a Informačných Technológií

To whom it concerns

Uchádzač: Marián Šimko
Názov: Modelovanie a extrakcia informácií pre inteligentný (slovenský) web

Predkladaná práca (zhrnujúca viacero prác uchádzača so spoluautormi) sa zaoberá veľmi aktuálnou problematikou počítačového spracovania informácií z webu. Ide o na výsosť dôležitú problematiku. V podstate, ide hlavne o počítačové spracovanie prirodzeného jazyka (slovenčiny) – ktorého súčasná úroveň je nedostačujúca a práca predstavuje vítaný príspevok.

Jadro habilitačnej práce je rozdelené do troch nadväzujúcich častí: modely informácií, spracovanie prirodzeného jazyka a samotná kapitola o extrakcii informácií z textu (obsah webu je väčšinou v texte). Každá z nich je ukončená „Sumárom príspevku autora do oblasti".

V kapitole o modeloch (popise konceptov a vzťahov medzi nimi) autor postupuje bez použitia formálnych modelov deskripčných logík a štandardov W3C. Používa knižku Cimiano 2006. Definícia1 (ontológia) a Definícia2 (ľahký doménový model) asi viac vyhovuje doménovému modelu vzdelávacieho systému Alef, ktorý používajú na experimenty. V paragrafe 3.4 síce spomína model OWL konzorcia W3, bolo by zaujímavé dozvedieť sa viac o prepojení týchto pohľadov.

V kapitole o spracovaní prirodzeného jazyka autor poskytuje veľmi dobrý prehľad problematiky. Pomocou experimentálneho prístupu otestovali s tímom ktorý vedie širokú škálu problémov a vyvinuli nástroje sprístupnené na portáli text.fiit.stuba.sk. Prekvapujúce je pre mňa tvrdenie, že použitie riešenia pre češtinu nedáva v žiadnom prípade uspokojujúcu úspešnosť pre slovenčinu. Záleží, koľko práce sa do toho vloží. Z vlastnej skúsenosti viem (keď som prekladal CZ ← → SK), že išlo väčšinou o jednoduché pravidlá (ů ← →u/ou, ě ← →e, ...). Nemalo by to byť príliš časovo náročné. Dokonca, viacerí Slováci, ktorí spolupracovali na UFAL MFF UK nástrojoch, urobili kroky aj pre spracovanie slovenčiny (bližšie info nemám). Bolo by tiež zaujímavé dozvedieť sa, ako sú modely z kapitoly3 prepojené s formalizmom kapitoly 4. Pre počítačové spracovanie je totiž dôležité mať ustálenú formalizáciu a reprezentáciu informácií.

Z hľadiska cieľov je najzaujímavejšia kapitola 5 o extrakcii informácií z textu. Autor považuje za najdôležitejšiu úlohu identifikácie a definície doménových konceptov. Nejde tu teda o spracovanie textu v zmysle NLP ale skôr o budovanie ontológií (ich poloautomatické dolovanie). Tvorba (aspoň trénovacieho vzorku) je poloautomatická a sleduje sa primeranosť záťaže pre používateľa. Veľmi sľubné je tu použitie sociálnych aspektov webu (analýza

používateľmi vytváraného obsahu pre tvorbu ľahkej sémantiky (Móro et al., 2011; Svrček, Šimko, 2014), metódy využívajúce sociálne aspekty pre extrakciu relevantných termov (Uherčík et al., 2013; Harinek, Šimko, 2013)).

Osobne mi trošku chýbala zmienka o dolovaní inštancií a spracovaní základnej štruktúry textu a to je vety (RDF trojica = holá veta). V citovaných prácach a aj na portáli text.fiit nástroje sú a zrejme kolektív previedol aj experimenty. Nástroj Synpar - Webová služba na syntaktickú analýzu mi spočiatku fungoval a potom som požiadal pár lingvistov aby mi ho otestovali a možno som ho nechtiac priviedol k pádu.

Táto habilitácia neposkytuje úplný prehľad autorových výsledkov (má toho oveľa viac) - tu sa viac sa sústreďuje na konceptualizáciu a tvorbu ontológií. Privítal by som aj popis netriviálneho prípadu použitia (use-case).

Samotná habilitačná práca je dobrým prehľadom problematiky a mapuje na hlavné publikované výsledky uchádzača. V tých je vidno obrovské množstvo práce tímu ktorý viedol. Je veľmi dobré, že vývoj došiel až do tohoto štádia – tím STU má totiž k dispozícii veľké množstvo experimentálnych prototypov a je tým dobre pripravený na ďalší výskum, školenie doktorandov.

Publikačné aktivity (špičkové i medzinárodné) i citácie sú veľmi dobré.

Práca je cenným vkladom do automatizácie spracovania slovenských textov na webe a dúfam, že podnieti aktivity aj iných kolektívov. Možno by mohol tím vedený ing. Šimkom pomýšľať aj na tvorbu testovacích sád na extrakciu informácií pre slovenčinu.

Prácu jednoznačne odporúčam prijať ako habilitačnú a po úspešnej obhajobe udeliť titul docent.

V Prahe 21.1.2017

Prof. Peter Vojtáš, DrSc.