Posudek vedoucího na diplomovou práci
Jana Dědka

Sémantická anotace dat z webovských zdrojů

Práce se věnuje poměrně náročné oblasti obohacování webovských zdrojů o počítačově srozumitelný význam. Konkétní analýza, návrh, prototypová implementace a experimenty v této práci se týkají její nejtěžší části a to oblasti dodatečné automatické anotace zdrojů psaných v českém jazyce. Je to důležitá součást vývoje v informatice, který směřuje k zpracování informací z webu bez lidského zásahu a má za cíl usnadnit zpracování programy které byly-budou napsány nezávisle od zdrojů. Je to problém na předním frontu výzkumu v softwarovém a datovém inženýrství.

Uchazeč prokázal že pochopil problém do hloubky. Prostudoval množství literatury, experimentoval s mnoha software jiných autorů (pro ne-české texty). Stavět na těchto řešeních se ukázalo neschůdné. V češtině máme naštěstí nástroje pro lingvistickou anotaci textů uložených lokálně v předem daném formátě. I když se jedná o nepropojené nástroje, se diplomand rozhodl použít nakonec tyto nástroje z pražské školy komputační lingvistiky (za to i příslušné konzultace jim patří i můj dík). Navíc použil vzdálený přístup k brněnské službě poskytující synonyma, hononyma a hierarchie v češtině. Autor dostatečně komentoval všechny peripetie rozhodování, zvažování alternativ a zdůvodnil své rozhodnutí.

Text práce najdřív zahrnuje popis deskripčních logik (DL) a W3C jazyků pro reprezentaci znalostí (RDF, DRF Schema, OWL). Pokračuje přehledem jazyků, nástrojů, projektů a uznávaných ontologií pro semantickou anotaci. Pokrývá taky lingvistickou anotaci a WordNet. Problematika jde do veliké šířky a zpracování je někdy až příliš stručné. Veliká stručnost postihuje i některé části o experimentech (detaily mnohých funkcí jsou komentovány jenom v zdrojových textech).

Nicméně můžu konstatovat, že uchazeč dokumentoval celý postup vypracování prototypového řešení a vysvětlil a zdůvodnit důležitá rozhodnutí učiněná při analýze a návrhu řešení (i s možnými alternativami a jejich důsledky).

Autor je až přehnaně skromný a velice často svoje řešení popisuje jako „pár řádek skriptu“. Ale celý postup, od webu až k extrahovaným datům, s propojováním různorodých nástrojů (v různých jazycích a složité struktuře) – i když často „jen pár řádek skriptu“ (ale někdy i obsáhlejší kód např. exprakční makro pro btred nebo program který ve WordNet-u hledá příbuzná slova) dáva výsledek, který umožnil netriviální náhled do problematiky. Teď už se dá pracovat na softwareovém řešení (asi nezanedbatelný počet člověko roků).

Kromě výše zmíněného, další věc kterou osobně vytýkám diplomandu je že neuvedl příklad ontologie a instancí obsahující vydolovaná data. Kromě své přehnané skromnosti je autor na sebe až příliš náročný, když píše, že chtěl ontologii „pro všechny problémy...a bude ve většině příkladů vyhovovat" (a takovou nenalezl). Osobně si myslím že některé makra pro dotazovací jazyk můžou sloužit jako příklady jednoduchých ontologií které budou jednou součásti „dokonalejších" ontologií (makra definují pohled jako koncept ve smyslu DL a OWL a např. používá pojmy action_type, injury_manner, participant, quantity viz str. 88 a 89). Je opravdu škoda, že tak malý krok autor neudělal. Každý výsledek odpovědi na tento dotaz je instancí této ontologie. Na druhé straně si cením práci jako prvvní prototyp v oblasti anotace českých webovských zdrojů, byl vyzkoušen ve dvou doménách, obsahuje extrakční pravidla a umožňuje použít i WordNet (který ale ve zmíněných aplikačních doménách je dosti chudý). Pro budoucí vývoj extrakce informací z lingvisticky anotovaných textů je cenný návrh oddělené interpretace dotazovacího jazyka Netgraph (str.89).

Ve výstupních datech sa občas objevuje chybové hlášení, např.
V `czsem\src\perl\output.xml` hlásí „`Cannot view XML input using XSL style sheet. Please correct the error and then click the Refresh button, or try again later. End tag 'SPAN' does not match the start tag 'sentece'. Error processing resource file:///C:/software/dedek_DP/czsem/src/pe...`
`<action type="zranit"><sentece>K dopravní nehodě dodávkového automobilu u Vysoké Libyně, ve směru na Jesenici, v...`
`:-2em">- <action type="přežít">  <sentece> Cyklistka může hovořit o skutečném štěstí, že přežila střet s nákladním vozem bez jediného škrábnutí.</sentece>`"

Závěr: Práci doporučuji k obhajobě, na které by autor měl odpovědět na výše zmíněné připomínky. Prosím dodat podrobnější diagram z obr. 6.1 ze str. 73 se znázorněním vlastních (i když jednoduchých (složitější neznamená nutně lepší)) kroků (i přechody ručnou úpravu nebo ručním spuštěním).

V Praze dne 31. 8. 2007

Prof. RNDr. P. Vojtáš, DrSc.
Vedoucí DP
KSI MFF UK