Posudek doktorské disertační práce Ing. Jaroslava Dytrycha „Sémantické anotování textu"

Předložená disertační práce se věnuje jednomu z aktuálních témat současné aplikované informatiky – anotování textů sémantickými metadaty, s důrazem na ruční a poloautomatické přístupy a na strukturované anotace. Za hlavní originální přínosy práce k výzkumu v oblasti považuji:

- Implementovaný *anotační systém* jako softwarový artefakt, a související datové artefakty: *komunikační protokol* a *datový formát* pro anotace
- *Uživatelské experimenty* provedené s pomocí tohoto systému (včetně srovnání s „konkurenčními" systémy).

Oba tyto vzájemně provázané „pilíře" disertační práce jsou rozsáhlé a propracované. V podrobné části posudku se zaměřím ve větší míře na rozbor anotačních (a ontologicko-modelovacích) experimentů než na anotační systém jako takový, protože jsem odborníkem na znalostní, nikoliv softwarové inženýrství. Anotační systém prošel praktickou evaluací charakteru případových studií v projektu EU DECIPHER (kde je doktorand spoluautorem dvou deliverables), a zřejmě i v novém projektu MixedEmotions. Vedle toho proběhla jeho experimentální evaluace srovnáním s několika obdobnými systémy (jak v kontextu anotování, tak i konstrukce ontologie jako souběžného procesu). Experimentální výsledky jsou z mého pohledu hodnotné, zejména pokud jde o experimenty označované autorem jako „pokročilé". Jejich rozsah je adekvátní disertační práci, a získané vhledy mohou být pro vývoj nových rozhraní velmi cenné. Škoda, že některé detaily experimentálního protokolu (který je zachycen jen formou zmínek roztroušených ve volném textu) nejsou uvedeny – podrobněji viz detailní poznámky níže.

Doplňkovým přínosem práce je provedená *komparativní rešerše* velkého počtu anotačních systémů.

Slabší stránkou práce z věcného hlediska je absence byť jen základní popisné *formalizace* navržených metod. Navíc i verbální popis *pokročilých anotačních metod*, jako je odlišení odkazované vs. vnořené anotace, generování nabídek anotací, sémantické filtrování hodnot atributů nebo generování „zjednoznačňovacího atributu", je velmi lakonický, chybí rozbor používaných algoritmů i podrobnější ilustrace příkladem. Čtenář je odkázán na intuitivní, generické pochopení podstaty metod ze zmínek roztroušených na více místech.

Dále se věnuji pojetí a obsahu jednotlivých kapitol, resp. jejich sekcí. Souhrnně bych k „toku výkladu" kriticky poznamenal, že stejný pojem, metoda nebo komponenta jsou často zmiňovány nebo vysvětlovány na více místech práce, navíc bez toho, že by jedno místo přímo odkazovalo na druhé.

- Rozhodnutí detailně popsat cíle práce v samostatné kapitole 2 považuji za správné. Trochu zvláštní ovšem je, že do této kapitoly je zařazena i dost velká část textu odpovídající spíše „stavu poznání v oblasti", a naopak následující kapitola 3 „Stav poznání v oblasti tématu práce" kromě jiného zavádí základní *pojmy* oboru – ty by měly být uvedeny spíše samostatně buď jako součást úvodní kapitoly, nebo hned za ní (jako tzv. „preliminaries").
- Překvapující je, že „existující řešení pro anotování" (sekce 3.3) neodpovídá výběru anotačních nástrojů, které byly v Příloze B zařazeny mezi „pokročilé". Očekával bych, že z rešerše vyplyne, které systémy jsou pro systém 4A „nejvážnějšími konkurenty", ty budou v kapitole věnované „stavu poznání" relativně podrobně rozebrány, a následně s nimi bude provedeno srovnání funkčností a podle možnosti i experimentální porovnání s vyvinutým systémem. V sekci 3.3 je

však uveden např. systém Annozilla, který charakterizuje „nemožnost tvorby strukturovaných anotací" a další omezení. Naopak chybí nástroje Pundit nebo MAT. Celkově postrádám jasnou metodiku výběru nástrojů pro výběr relevantních nástrojů.

- Termín „metodika" je použit v sekci 4.1, v souvislosti s přístupem ke srovnávání nástrojů, ovšem poněkud nadsazeně, jde spíše o popis sledovaných kritérií ve volném textu, místy poněkud chaotický. Náznak metodiky je snad jen v posledních třech odstavcích sekce, ovšem postup je popsán jen vágně. Mluví se o „několika vybraných nástrojích", aniž by bylo explicitně řečeno, co je podmínkou výběru.
- Sekce 4.2 se věnuje požadavkům na anotační systém. Je trochu překvapivé, že autor pouze přebírá dva seznamy požadavků z literatury, aniž by je nějak doplnil, případně kriticky přehodnotil.
- Část 5, popisující samotný návrh anotačního systému, působí přesvědčivě, jak jsem ale zmínil, nejsem schopen kvalitu softwarového řešení (zejména serverové části) kompetentně zhodnotit. Možná ne zcela šťastné je promíchání popisu implementace se zmínkami o zpětné vazbě od uživatelů v projektu DECIPHER.
- „Případy použití" v sekci 5.7 jsou poněkud nesourodé. Zatímco podsekce 5.7.1 se věnuje spíše hypotetickému generickému použití v oblasti publikování, podsekce 5.7.2 popisuje konkrétní případ užití v rámci projektu DECIPHER. Podsekce 5.7.3 („Rozšiřování ontologie") je rozsáhlá a detailní, její velká část však vlastně o rozšiřování ontologie nemluví, tohoto tématu se týká jen část začínající na konci s. 59.
- Pokud jde o sekci 6 věnované experimentům, soustředím se na „pokročilé" experimenty, protože „testování prvotního prototypu" je již staršího data, zaznamenané výsledky jsou jen „kvalitativního" charakteru, a jejich popis je dost kusý; v případě nasazení v projektu DECIPHER pak zřejmě nešlo o cílené experimenty navržené autorem disertace, ale spíše o různorodé využívání projektovými partnery mimo možnost přímé kontroly (?).
- Srovnání s GATE a RDFaCE v sekci 6.4.1 se zdá přesvědčivé; vypovídací schopnost výsledků (rozdílů mezi systémy) však může záviset na celkovém počtu anotovaných instancí, který v textu není uveden. Podobně jasná je preference zjednoznačňujícího atributu v případě ověřování optimálního rozsahu zobrazovaných informací v sekci 6.4.2 (pokud dobře počítám, ze 720 případů bylo u této varianty 15 chyb, zatímco u varianty s detailními informacemi 45 – mimochodem, nevím, proč zde i jinde nejsou v tabulce i absolutní čísla). Rozdíl mezi dvěma způsoby zobrazení alternativních nabídek anotací (sekce 6.4.3) už tak průkazný není, a těžko z něj lze dělat závěry. Naopak, přínos sémantického filtrování byl v sekcích 6.4.4 a 6.5.1 prokázán jasně. Sekce 6.5.2 poskytuje zajímavý vhled do vzorů interakce z hlediska rozbalování/sbalování nabídek. Konečně, scénář testování rozšiřování ontologie v sekci 6.6 je nápaditý; určitým omezením ovšem je, že zřejmě neřeší otázku kvality vytvořené ontologie. Není patrné, zda tvorba ontologie přes 4A nevede k nekoherenci ontologie nebo neomezuje tvorbu určitého typu struktur. V práci není jasně deklarováno, že by tvorba struktury ontologie pomocí anotování měla být vhodná jen pro ontologie, které budou následně používány opět jen pro anotování textů a nikoliv jako prostředek logického odvozování. Náznak takového předpokladu je v sekci 7.7: „Navržené řešení předpokládá spolupráci přímo nad texty, při jejichž analýze má být výsledná ontologie využita"; proti tomu však mluví text na konci sekce 2.2: „Při tom by mělo být ukázáno, že strukturované anotace umožňují lepší popis sémantiky a snazší strojové zpracování anotací (např. podpora automatického odvozování informací)."
- Kapitola 7 je velmi kvalitně zpracována (předpokládám, že byla psána v pokročilé fázi studia, zatímco některé jiné části byly spíše adaptovány, občas ne zcela důsledně, ze starších textů).
- Tabulky v přílohách A a B odrážejí velký objem vložené práce. Tato práce ovšem nebyla zcela zúročena tím, že by tabulky byly důkladněji rozebrány v textu. Celý rozbor zahrnuje jen o něco

víc než stranu, v sekci 7.3, tj. téměř na konci práce. Jak už bylo naznačeno, patřil by do kapitoly věnované stavu poznání (3.3), a to v rozvinutější podobě, např. včetně explicitní interpretace „hodnot s otazníky" z tabulek (které budí dojem spíše pracovních poznámek než rigorózního srovnání) – pak by srovnání systémů bývalo mohlo být plnohodnotným třetím „pilířem" disertace vedle softwarových/datových artefaktů a realizovaných experimentů.

- Je překvapivé, že navzdory velkému rozsahu příloh mezi nimi chybí text instrukcí předaných testerům před experimenty (např. jakou formou byl požadován „určitý kompromis mezi časem stráveným na jedné anotaci a výslednou kvalitou"?), dotazníků vyplňovaných po experimentu, i kompletní agregované výsledky dotazníků.

Kontextem, ve kterém probíhala významná část výzkumu, je *projekt EU DECIPHER*. Zmínky o něm jsou roztroušeny na mnoha místech disertace (mimochodem, s nejednotnou kapitalizací), býval by si ale zasloužil samostatnou sekci se stručným popisem. Podobně by býval měl být popsán i nový projekt MixedEmotions.

Za velmi pozitivní považuji úsilí o spolupráci se *skupinou W3C* zaměřenou na předmětnou oblast. Skutečnost, že systém 4A mohl v projektu DECIPHER úspěšně absorbovat nezávisle vzniklý formát OpenAnnotation, bez zásadní změny aplikace a komunikačního protokolu, je dobrým indikátorem jeho flexibility.

Pokud jde o *publikační činnost*, výsledky disertačního projektu byly postupně publikovány ve čtyřech sbornících z konferencí/workshopů. Přestože chybí skutečně „reprezentativní" publikace (v hlavní sekci některé vysoce selektivní konference nebo v prestižním mezinárodním časopise), považuji publikační výsledky za přijatelné pro obhajitelnost disertační práce v oboru podle tuzemských měřítek. Nejvýznamnější publikací je (na webu k dnešnímu datu zatím nedohledatelný, předpokládám ale, že přijetí je doloženo) dlouhý příspěvek ve výběrovém post-konferenčním sborníku z ICAART 2016, který je/bude vydán v řadě Springer LNCS. Doktorand je u tohoto příspěvku (i u jeho kratší verze prezentované na ICAART) prvním autorem. Také obě dřívější (druhoautorské) publikace byly umístěny na fóra v odborné komunitě respektovaná - konferenci FLAIRS 2015 a workshop SePublica 2011 konaný při evropské konferenci o sémantickém webu, ESWC. *Citace* v recenzovaných publikacích se mi dohledat nepodařilo; na Google Scholar jsou pouze dva odkazy na příspěvek ze SePublica, a to z disertační práce J. Corneliho (kolegy z partnerského pracoviště projektu EU DECIPHER). Lze ale předpokládat, že novější publikace obsahující pokročilé výsledky si již své citace najdou.

Kvalitu *textového zpracování* disertační práce bych hodnotil jako průměrnou. Na jedné straně je text psán s přiměřenou mírou pečlivosti, množství překlepů je přijatelné a nenarazil jsem na výrazné typografické prohřešky. Na druhé straně lze vytknout nedostatečné užívání prostředků strukturace např. číslovaných seznamů (trochu paradoxně vzhledem k vyzdvihování výhod strukturovaných anotací v práci samotné). Text je z podstatné části psán formou odstavců volného textu, navzdory tomu, že se v něm houfně vyskytují různé výčty. Obzvláště markantní to je v části o experimentech, kde se mluví o různých „obdobích", „částech", „sériích" a „sadách" experimentů, aniž by hierarchická struktura celé soustavy experimentů byla někde přehledně vizualizována.

Tabulky a obrázky jsou zařazeny v dostatečném rozsahu a vhodným způsobem, zejména u obrázků (diagramy, např. 4.1, i většiny snímků obrazovek) ovšem postrádám detailnější vysvětlení jejich detailů – odkaz z textu zpravidla jen souhrnně uvádí, že architektura nebo funkčnost XY je vidět na Obr. Z.

Seznam literatury je přiměřeně rozsáhlý, použití číselných odkazů však nepovažuji za ideální (u delších textů typu disertací, kdy se často na stejný zdroj odkazuje opakovaně, bych upřednostnil harvardský styl citací). Navíc jsou čísla odkazů kvůli „odkazové" změně barvy v černobílém tisku prakticky nečitelná. V práci také postrádám seznam obrázků a tabulek, popřípadě glosář nebo rejstřík termínů.

Volba *českého jazyka* je u informatické disertace (dokonce navázané na projekty EU) podle mých zkušeností méně obvyklá, ale nepovažuji ji za závadu, může mít naopak i pozitivní efekt z hlediska etablování české terminologie oboru. Nutno ovšem poznamenat, že v disertaci se občas s pojmy pracuje „volněji", viz některé detailní připomínky níže. Dále, některé diagramy byly ponechány v angličtině – převod do češtiny by měl být důslednější.

Drobné detailní připomínky, které jsem si při čtení poznamenal:

- S. 11: „Uživatelé mohou označit stejnou věc vzájemně se vylučujícími termíny. Pokud je jeden termín označen více protichůdnými tagy, ..." Zde jde zjevně o míchání pojmů „termín" a „tag".
- S. 12: „Pokud lidé označí nějaký pojem tagem, lze tento pojem chápat jako entitu..." Výraz „pojem" se ve znalostním inženýrství zpravidla neztotožňuje s výskytem jazykového výrazu, který lze v textu anotovat. „Pojem" je spíše ontologickou entitou (zpravidla typem spíše než instancí), která se při anotování textu přiřazuje.
- S. 11 a 12: termín „emerging semantics" se zde překládá dvěma různými českými výrazy, „objevující se sémantika" a „nově objevená sémantika".
- S. 40: „atributy, které nepatří k žádnému typu (v ontologii vztahy bez subjektu)." Tato formulace v terminologii RDF/OWL nedává smysl. Každá trojice RDF má subjekt. Zjevně jsou míněny vztahy (vlastnosti), které v ontologii nemají uveden definiční obor (domain). Obdobně i pro další obraty použité v tomto odstavci. Naopak na s. 45 je použit v tomto kontextu nepříliš zavedený termín „podmět" (běžně se v češtině používá „subjekt"): „Formát anotace je založen na RDF/XML [11], kde podmětem je vždy anotace."
- S. 40: „nastavit typ hodnoty, který může být nejen typem atributu, ale i jeho podtypem." Pojem „podtyp atributu" nebyl do této chvíle nikde zmíněn. Jedná se zřejmě o zúžení typu atributu pro konkrétní případ jeho použití? I zde, stejně jako na mnoha jiných místech, citelně chybí ilustrace na příkladech.
- S. 45: Nejsem přesvědčen, že rozšíření RDF/XML na reprezentaci n-tic je vhodné řešení. Interoperabilita s nástroji pro RDF se tím ztrácí, jedná se pak jen o nadbytečně komplikované struktury XML neodpovídající žádnému standardu „vyšší úrovně".
- S. 46: „Skupina uživatelů v linearizovaném názvu není, neboť si ji uživatel zvolí v nastavení a nemá smysl, aby ji opakovaně zadával." Tato věta v daném kontextu není jasná. Proč má být v názvu skupina uživatelů?
- S. 60: „Fragmenty textu anotované tímto způsobem mohou být převzaty jako příklady použití..." Jak si autor představuje zařazení příkladů použití do ontologie? Jako hodnoty nějaké textové anotační vlastnosti? Které?
- S. 60: „nástroj podporuje import a export komplexních ontologií v OWL" O jakou komplexitu z hlediska OWL se jedná? Lze ji vyjádřit ve smyslu expresivity určité deskripční logiky, zavedených profilů OWL, npod.?
- S. 63: „Z těchto dat je extrahován prostý text, který je následně vertikalizován" Co to znamená?
- S. 72: „První experiment, srovnávající anotační systémy, spočíval v anotaci událostí." Jestli to dobře chápu, podobně byl zaměřen i experiment se sémantickým filtrováním?
- S. 77: „Celkový vysoký počet chyb (ve sloupci „Nesprávné hodnoty") lze vysvětlit velmi striktním porovnáním s referenčními anotacemi... Někteří z nich však zadali hodnotu 1950" Zamítnutí této hodnoty v případě textu „žena ve svých padesáti letech" mi nepřipadá nijak striktní, jde o něco sémanticky zcela jiného.
- S. 81: „Byly připraveny dvě datové sady po 15 úryvcích textu zmiňujících vybrané entity, kde 3 z nich (20 %) v každé sadě nebyly pokryty referenčními zdroji" Z textu není jasné, zda uživatelé o této proporci věděli nebo ne. To mohlo mít na výsledky experimentu vliv.

- S. 86: „Aby bylo možné korektně pracovat s různým počtem alternativ a různými pozicemi správné alternativy v seznamu nabídek pro jednotlivé zmínky o entitách v textu, naměřený celkový čas byl vždy vydělen pořadím správné alternativy." Tato modifikace mi připadá neúměrně silná. Předjímá, že uživatel nemá nikdy potřebu po nalezení předpokládaného správného řešení ještě procházet ta další (přitom autor sám uvádí, že toto je pozorováno jen v 86% případů). Dovedu si představit i možnost, že vyloučení alternativy zabírá kratší čas než její spolehlivě potvrzení. Z toho důvodu by stálo za to zvážit možnost např. v nějakém dalším experimentu sledovat pohyb oka.
- S. 87: „koresponduje s reálnou situací, kdy klient spolupracuje se znalostním inženýrem" Sice to z kontextu je jasné, ale přesto by bylo dobré upozornit, že termín „klient" se v této sekci začíná používat v jiném smyslu, než v celé předchozí části práce.

U obhajoby bych rád slyšel odpovědi na následující *doplňující otázky*:

- Jak náročné by bylo integrovat 4A, formou editorového doplňku, s některým běžným CMS, např. WordPressem?
- V textu není jasně demonstrováno, jak se v případě obohacování ontologie začlení nová entita do její struktury. Jak se např. rozliší, jestli chtěl uživatel přidat třídu nebo instanci? Může být přes anotační rozhraní přidána i vlastnost? Prosím o vysvětlení a ukázání konkrétních příkladů. Návazně prosím i o stručné vyjádření, jaké typy logického odvozování jsou ontologií (resp. jejími částmi vzniklými přes anotační rozhraní) podporovány.
- Spíše formálně zaměřená otázka: jaký je podíl dalších spolupracovníků na jednotlivých výsledcích v disertaci prezentovaných? (Tento podíl bývá v disertacích zpravidla explicitně vymezen.)

Závěrem shrnuji, že disertační práce, navzdory některým připomínkám k jejímu zpracování, podle mého názoru splňuje kritéria originální vědecké práce v oboru, obsahuje výsledky zajímavé v mezinárodním měřítku, a v případě úspěšného průběhu obhajoby proto **doporučuji** uchazeči udělit titul doktor.

V Praze, dne 7.1.2017

doc. Ing. Vojtěch Svátek, Dr.
katedra informačního a znalostního inženýrství
fakulta informatiky a statistiky Vysoké školy ekonomická v Praze