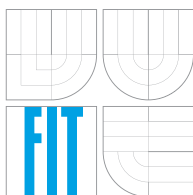




VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS
AND MULTIMEDIA

SÉMANTICKÁ ANOTACE TEXTU

SEMANTIC ANNOTATION OF TEXT

TEZE DISERTAČNÍ PRÁCE - AUTOREFERÁT
PHD THESIS SUMMARY

AUTOR PRÁCE
AUTHOR

Ing. JAROSLAV DYTRYCH

ŠKOLITEL
SUPERVISOR

Doc. RNDr. PAVEL SMRŽ, Ph.D.

BRNO 2016

Abstrakt

Tato práce se zabývá inteligentními systémy pro podporu sémantického anotování textu. Rozebírá motivaci vzniku takových systémů a stav poznání v oblastech souvisejících s jejich používáním. Popisuje také nově navržený a implementovaný anotační systém, realizující unikátním způsobem pokročilé funkce sémantického filtrování a prezentace alternativ anotací. Výsledky provedených experimentů jasně ukazují výhody zpracovaného řešení. Demonstrují také, že uživatelského rozhraní anotačních nástrojů významně ovlivňuje proces anotování. Následně je provedena optimalizace zobrazovaných informací pro úlohu desambiguace mnohoznačných jmen a jsou experimentálně vyhodnoceny přínosy metod navržených pro zvýšení rychlosti anotování a kvality vytvářených anotací. Srovnáním s obecným nástrojem Protégé je rovněž ukázán přínos vytvořeného systému pro kolaborativní tvorbu ontologií, které mají být “ukotveny” v textu. Na závěr jsou analyzovány a shrnuty veškeré dosažené výsledky.

Abstract

This thesis deals with intelligent systems for support of the semantic annotation of text. It discusses the motivation for creation of such systems and state of the art in the areas of their usage. The thesis also describes newly proposed and realised annotation system which realizes advanced functions of semantic filtering and presentation of annotation suggestion alternatives in a unique way. The results of finished experiments clearly show the advantages of proposed solution. They also prove that the user interface of the annotation tools affects the annotation process. The optimisation of displayed information for the task of disambiguation of ambiguous entity names was done and proposed methods to speedup and increase of quality of the created annotations was experimentally evaluated. The comparison with the Protégé general tool has proven the benefits of created system for collaborative ontology

creation which should be anchored in the text. In the conclusion, all achieved results are analysed and summarized.

Klíčová slova

sémantické anotace, strukturované anotace, protokol pro přenos anotací, srovnání anotačních nástrojů, W3C Web Annotation, W3C Open Annotation, nabízení anotací, alternativní nabídky anotací, sémantické filtrování, vytváření ontologií na základě textu

Keywords

semantic annotations, structured annotations, protocol for annotation transmission, comparison of annotation tools, W3C Web Annotation, W3C Open Annotation, annotation suggestions, alternative suggestions of annotations, semantic filtering, ontology creation based on text

Citace

DYTRYCH, Jaroslav. *Sémantická anotace textu*. Brno, 2016. Teze disertační práce - autoreferát. Vysoké učení technické v Brně, Fakulta informačních technologií. Školitel Pavel Smrž.

Sémantická anotace textu

Poděkování

Děkuji Doc. RNDr. Pavlu Smržovi, Ph.D., za odborné vedení, vstřícnost a ochotu při řešení této práce. Děkuji také své rodině za podporu a trpělivost.

© Jaroslav Dytrych, 2016.

Tato práce vznikla jako školní dílo na Vysokém učení technickém v Brně, Fakultě informačních technologií. Práce je chráněna autorským zákonem a její užití bez udělení oprávnění autorem je nezákonné, s výjimkou zákonem definovaných případů.

Obsah

1 Úvod	3
2 Cíle práce	7
2.1 Srovnání uživatelských rozhraní anotačních nástrojů	7
2.2 Vytváření strukturovaných anotací	8
2.3 Anotování kdekoliv a kdykoliv	9
2.4 Zvýšení kvality vytvářených anotací	10
2.5 Formalizace znalostí na základě anotování textu . .	11
3 Stav poznání v oblasti tématu práce	13
3.1 Definice pojmu anotace	13
3.2 Existující řešení pro anotování	14
4 Zvolené metody řešení	17
5 Realizovaný anotační systém	21
6 Provedené experimenty	27
7 Dosažené výsledky a jejich analýza	37
7.1 Protokol pro přenos anotací a formát anotace . . .	38
7.2 Vytvořený anotační systém	39
7.3 Vyhodnocení srovnání anotačních nástrojů	40
7.4 Určení optimálního množství zobrazovaných infor- mací	43

7.5	Přínos alternativních nabídek anotací	45
7.6	Přínos sémantického filtrování	47
7.7	Kolaborativní vytváření ontologie	48
7.8	Souhrnné výsledky	49
8	Závěr	51
	Literatura	57
	Přílohy	65
	Seznam příloh	67
A	Strukturovaný životopis	69
B	Seznam publikací	73
C	Seznam inženýrských děl	75

Kapitola 1

Úvod

I přes velkou snahu o formalizování znalostí je stále většina textových informací dostupná pouze ve formě textu v přirozeném jazyce. Dosáhnout strojového porozumění takovému textu je velmi obtížné, což mimo jiné ztěžuje realizaci původní vize sémantického webu. Nejmodernější nástroje pro extrakci informací z textu jsou omezené pouze na úzké sady případů, pro které byly trénovány. V ostatních případech je stále potřebné manuální či poloautomatické vytváření metadat. V této práci se proto zaměřuji na jeden z přirozených způsobů formalizace znalostí – anotování textu.

Existuje celá řada anotačních nástrojů umožňujících vytváření jednoduchých textových poznámek, případně odkazů na referenční zdroje popisující zmíněné entity. Pokročilejší nástroje umožňují i strukturování anotací pomocí odkazů mezi nimi či pomocí atributů. Tato funkcionalita je však podporována pouze v omezené míře. Navíc jsou příslušné nástroje často úzce zaměřeny na konkrétní případy užití a pro obecnější úlohy je nelze použít vůbec, případně jen s mnohými omezeními.

Studii, které se zabývají srovnáním nástrojů pro manuální či poloautomatické anotování textu z uživatelského hlediska a možnostmi jejich využití, existuje překvapivě málo. Většina srovnání se totiž zaměřuje na nástroje pro automatické generování nabí-

dek anotací v pozadí, které je však u řady moderních nástrojů možné jednoduše nahradit a pro určení možností samotných anotačních nástrojů je tak toto srovnání irelevantní. Existující studie zaměřené na uživatelské rozhraní a využitelnost nástrojů jsou navíc často zastaralé a neposkytují dostatek informací pro určení možností strukturování anotací a nasazení nástrojů při různých úlohách. V rámci svojí práce jsem proto připravil novou studii srovnávající více než 60 nástrojů. Vybrané pokročilé nástroje pak byly detailněji srovnány z hlediska možností nasazení pro komplexní anotační úlohy a 3 z nejlepších byly srovnány v praktických experimentech.

Současná podpora pro zjednodušování entit a vytváření strukturovaných anotací komplexních vztahů v textu byla shledána nedostatečnou. Většina nástrojů pro zjednodušování nabízí pohled prezentující pouze typ entity a adresu stránky v referenčním zdroji, která ji popisuje. Při vytváření vztahů je pak většina práce ponechána na uživateli, pro kterého se často jedná o netriviální úlohu. S tím souvisí vysoká náročnost anotování pro uživatele i potenciální vysoký počet chyb při delším anotování. Ve svojí práci se proto zabývám návrhem a ověřením nových konceptů uživatelských rozhraní anotačních nástrojů umožňujících snížení náročnosti anotování pro uživatele, a tím i zrychlení procesu anotování a zvýšení kvality vytvářených anotací. Mezi nejdůležitější z nich patří sémantické filtrování a pokročilý způsob prezentace alternativních nabídek anotací.

Vzhledem k potřebě prakticky ověřit nové koncepty bylo nutné navrhnout a implementovat nový anotační nástroj. Nástrojů pro ověření nových konceptů pro sémantické anotování vzniklo v minulosti již několik. Jejich specializované zaměření však obvykle vedlo k tomu, že se vývoj soustředil jen na konkrétní vybrané aspekty uživatelských rozhraní a nástroje nebyly šířeji využitelné. Já jsem si naopak vytkl cíl vytvoření univerzálního modulárního nástroje se širokými možnostmi konfigurace, umožňujícího nejen ověření navržených konceptů, ale i praktické nasazení v reálném

prostředí a poskytnutí vhodné platformy pro budoucí výzkum v oblasti anotování textu. Vytvořený nástroj umožňuje snadné doplnění a úpravu funkcionality pro zkoumání vlivu různých aspektů uživatelského rozhraní na proces anotování i jednoduché vestavění a ověření zcela nových konceptů. Pro prokázání praktické využitelnosti byl nasazen i v evropském projektu DECIPHER.

S výslednou verzí vytvořeného anotačního systému pak byly provedeny experimenty, které demonstrují, že uživatelské rozhraní anotačních nástrojů významně ovlivňuje proces anotování. Bylo při nich prokázáno, že nově navržená funkcionalita sémantického filtrování a prezentace alternativ anotací umožňuje zvýšení rychlosti anotování a kvality vytvářených anotací. Srovnáním s obecným nástrojem Protégé je rovněž ukázán přínos vytvořeného systému pro kolaborativní tvorbu ontologií.

Kapitola 2

Cíle práce

Primárním cílem mé práce je ukázat, že propojení strukturování znalostí s tagováním, podpořeným pokročilými anotačními nástroji, zjednodušuje proces získávání a strukturování znalostí a na základě experimentů vyhodnotit a kvantifikovat přínosy tohoto postupu. Pro dosažení tohoto cíle se zabývám různými aspekty anotování se zaměřením na práci uživatelů. Tyto aspekty se vzájemně ovlivňují a jejich vhodná kombinace může zvýšit produktivitu uživatelů i kvalitu vytvářených anotací. Kvantifikace těchto zlepšení by měla patřit k hlavním přínosům mé práce.

2.1 Srovnání uživatelských rozhraní anotačních nástrojů

Existuje celá řada anotačních nástrojů, avšak jen malé množství studií se zabývá jejich srovnáním. Např. Diana Maynard [16] srovnává 6 nástrojů z hlediska různých aspektů, jakými jsou např. použitelnost (instalace, dokumentace, vzhled, ...), přístupnost (funkcionalita uživatelského rozhraní) a interoperabilita (formáty a platformy). Článek [15] pak uvádí srovnání z hlediska uživatele provádějícího manuální anotování, uživatele výsledných anotací, tvůrce

korpusů a vývojáře systémů, do kterých je anotační funkcionalitu potřeba zabudovat. Další srovnání, jako např. [33], se pak zabývají převážně technickými aspekty a perspektivu uživatele opomíjejí. V průběhu svojí práce jsem neobjevil žádnou studii, která by se detailně zabývala srovnáním nástrojů z hlediska uživatele a vyhodnocovala dopad jednotlivých částí uživatelského rozhraní na proces anotování.

Prvním dílčím cílem mojí práce tedy bylo srovnat jednotlivé nástroje z uživatelského hlediska a vyhodnotit, jak jednotlivé části uživatelského rozhraní ovlivňují činnost uživatele. Následně bylo cílem kvantifikovat, nakolik uživatelské rozhraní ovlivňuje rychlost anotování a kvalitu vytvářených anotací.

2.2 Vytváření strukturovaných anotací

Většina existujících nástrojů nabízí možnost anotování textovou poznámkou, případně tagy. Strukturované anotace jsou spíše výjimkou a ve většině nástrojů nejsou dostupné, nebo je jejich tvorba značně omezená či komplikovaná. Dokumenty, které se v průběhu anotování mění, jsou pro velké množství systémů rovněž nevyřešený problém.

Některé současné systémy dále umožňují kolaborativní práci s anotacemi, ale souběžný přístup obvykle není dobře ošetřen. Anotace jsou aktualizovány pouze v okamžiku načtení či obnovení dokumentu a následně může dojít k situaci, kdy někdo edituje neaktuální verzi anotace. To může v konečném důsledku vést i k fatální ztrátě dat. Většina systémů navíc pracuje pouze se statickými dokumenty či snímky dokumentů z určitého časového okamžiku. Pokud se dokument mění, může být problém s aktualizací anotované verze. Další komplikaci pak představuje uživatel, který si dokument pozmění lokálně u sebe a pak provádí synchronizaci s centrálně uloženou verzí. V této situaci nelze využít malé granularity změn a aktualizace pozic anotací může být netriviální.

Dalším cílem mojí práce bylo navrhnout systém, který umožní využití strukturovaných anotací spolu s klasickými poznámkami a tagováním. Mělo být podpořeno vytváření jak jednoduchých anotací, na které jsou uživatelé zvyklí z existujících nástrojů, tak i strukturovaných anotací s atributy různých datových typů, vnořenými anotacemi a vazbami na jiné anotace. Pomocí strukturování pak mělo být možné vyjádřit složitější relace mezi anotovanými pojmy. Součástí práce má být i řešení pro spolupráci v reálném čase.

Systém by měl být dostatečně obecný, aby jej bylo možné využít k různým účelům (včetně těch, které jsou uvedeny v návrhu konceptuálního frameworku pro digitální anotační systémy na webu od Niranatlamphonga, Choochaiwattana a Springa [19]) a dosáhnout tak i rozšíření možností využití anotací. Při tom by mělo být ukázáno, že strukturované anotace umožňují lepší popis sémantiky a snazší strojové zpracování anotací (např. podpora automatického odvozování informací), přičemž pokud tyto anotace využijeme při kolaborativním anotování v reálném čase, usnadní se tím spolupráce při tvorbě dokumentů, ontologií a dalších činnostech.

2.3 Anotování kdekoliv a kdykoliv

Dostupné nástroje umožňují anotovat dokumenty v určitém prostředí, kterým je obvykle webový prohlížeč nebo editor dokumentů. S tím je spojeno i omezení na anotované dokumenty, a to na takové formáty, které se v daném prostředí využívají. V žádném ze zkoumaných nástrojů není možné dokument anotovat ve více formátech současně (např. tvůrce webové stránky ji anotuje v textovém editoru ve formátu prostého textu či v syntaxi wiki, zatímco návštěvník této stránky ji anotuje ve webovém prohlížeči ve formátu HTML). Rovněž není možné anotovat v různých prostředích současně (např. ve webovém prohlížeči a v desktopovém textovém editoru).

Cílem práce bylo také vytvoření řešení, které umožní anotovat text v různých prostředích a formátech současně a převádět anotace mezi jednotlivými formáty dokumentu. Uživatel by měl mít možnost anotovat text v textovém editoru na svém pracovním počítači a pokračovat v jednoduchém editoru s webovým rozhraním na chytrém mobilním telefonu v průběhu konference. Anotovat tedy má být možné kdekoliv (v libovolném prostředí) a kdykoliv (kdykoliv má uživatel přístup k síti, sám nebo současně se spolu-pracovníky). Dosud nebyl k dispozici žádný otevřený protokol pro přenos anotací, který by splňoval náročné požadavky na kolabora-tivní anotování v heterogenním prostředí. Součástí práce tak musí být i návrh nového univerzálního protokolu pro přenos anotací a nového formátu anotace.

2.4 Zvýšení kvality vytvářených anotací

Při návrhu řešení je třeba dbát na to, aby byl výsledný systém co nejvíce konfigurovatelný a umožnil nastavovat množství zobra-zovaných informací a další aspekty uživatelského rozhraní. Srovná-ním různých nastavení takového systému je potom možné nalézt nejvhodnější konfiguraci, která umožňuje uživateli např. dosáh-nout za stejnou jednotku času vyšší kvality anotací.

Dalším dílčím cílem práce tedy bylo dosáhnout výše uvedené modularity vytvořeného systému. Součástí této části práce bylo také s využitím systému experimentálně určit, jaký vliv na proces anotování má množství informací, které jsou uživateli zobrazovány. Na základě výsledků je potom možné optimalizovat množství zob-razovaných informací tak, aby se zvýšila rychlost anotování a kva-lita vytvářených anotací.

V neposlední řadě bylo cílem kvantifikovat přínosy pokročilých funkcí využitelných při poloautomatickém anotování – sémantic-kého filtrování na základě šablon pro jednotlivé sémantické typy a pokročilé prezentace alternativních nabídek anotací. Praktic-kými experimenty by následně mělo být ukázáno, že tyto funkce

zvyšují rychlost anotování a kvalitu vytvářených strukturovaných anotací.

2.5 Formalizace znalostí na základě anotování textu

Vytváření ontologií a dalších prostředků formalizace znalostí představuje náročný proces [11]. Specializované nástroje pro tvorbu ontologií, jako např. WebProtege¹, často vyžadují netriviální znalost formálních aspektů reprezentace znalostí. Oboroví odborníci, pracující přímo s dokumenty v určité oblasti (např. pracovníci muzeí), však takové znalosti obvykle nemají. Proto je ontologie často vytvářena specialistou, který pomocí rozhovorů, různých dotazníků a poskytnutých materiálů získává informace od doménových odborníků. Přichází se tak o možnost vytváření ontologií přímo při práci s textem. Pevně dané ontologie v rychle se vyvíjejícím prostředí obvykle neobstojí. V reálném prostředí je třeba znalostní struktury rozšiřovat a aktualizovat. Pro správce ontologie však může být obtížné objevovat chybějící koncepty a jejich významy, pokud není tato činnost přímo propojena se zpracováním relevantních dokumentů. Bohužel jen málo existujících nástrojů umožňuje anotovat pomocí dříve vytvořených formalizovaných znalostních struktur a současně tyto struktury rozšiřovat.

Cílem práce je tedy umožnit jak kolaborativní vytváření ontologie při anotování, tak i snadné přidávání nových konceptů do existujících ontologií, při kterém bude mít jejich správce k dispozici řadu příkladů jejich využití, doložených v konkrétních textech. Následně je cílem ukázat, že kolaborativní rozšiřování ontologie s využitím reálných anotovaných dat usnadňuje proces tvorby a aktualizace formalizovaných znalostí.

¹<http://webprotege.stanford.edu>

Kapitola 3

Stav poznání v oblasti tématu práce

V této kapitole se zabývám současným stavem poznání v oblasti anotování textu. Nejprve se zaměřuji na samotný pojem anotace a následně na existující programová řešení z dané oblasti.

3.1 Definice pojmu anotace

Historicky je slovo anotace definováno jednak jako poznámka přidaná jako vysvětlení k nějaké literární práci a jednak jako akt samotného anotování. Více o historii slova anotace lze nalézt v [2]. Dále budu slovem anotace označovat doplňující informaci přidanou k textu bez omezení významu této informace či charakteru anotované části textu.

Důležitým pojmem je i tzv. tag (značka či štítek), což je krátká, obvykle jednoslovná, textová anotace. Tagging (tagování či značkování) je tedy specifický druh anotování, který spočívá v doplňování krátkých anotací k textu a je považován za nejjednodušší způsob přidávání metadat k fragmentům textu. Více lze nalézt v článku od Völkela a Hallera [30].

3.2 Existující řešení pro anotování

Pro anotování v prostředí webu existuje celá řada nástrojů, přičemž první z nich vznikly před více než 15 lety (viz článek LaLiberte a Bravermana [14] nebo např. GrAnT [24]). Pro ukázkou jsem zvolil několik nástrojů, u kterých jsou patrné obvyklé přístupy, možnosti anotování a sdílení anotací.

BRAT

BRAT [28] je obecný lingvistický editor anotací, který již byl využit k přípravě celé řady datových sad. Klíčovými charakteristikami tohoto nástroje jsou plochá struktura základních anotací a jednoduchý způsob jejich prezentace. Nástroj je vhodný pro vysoce specializované úlohy, jakými je např. vyhledání koreferencí nebo extrakce vztahů mezi biomedicínskými entitami s unikátními názvy. Odkazování do znalostní báze a hierarchické anotování komplexních vztahů, jakými jsou např. události, není podporováno.

RDFaCE

RDFaCE [13], [12] je rozšíření WYSIWYG editoru TinyMCE [18] napsaného v jazyce JavaScript. Tento editor je využíván v řadě systémů pro správu obsahu (CMS) a anotační rozšíření tak lze využít v mnoha prostředích na webu. Anotace jsou ukládány přímo do dokumentu ve formátu RDFa. Vytvářejí se označením anotované části textu a vyplněním formuláře s atributy anotace. Zobrazení je pak v podobě okénka se stručnými informacemi, které se zobrazí při najetí myši, a pro detailní informace je třeba opět otevřít editační formulář. Nevýhodami jsou nemožnost spolupráce v reálném čase, problémy s uchováním anotací při editaci textu a relativně pracné vyplňování formulářů při vytváření anotací.

GATE Developer

GATE Developer [10] je vývojářské prostředí, které poskytuje sadu nástrojů pro tvorbu programových komponent pro zpracování přirozeného jazyka. Jedním z integrovaných nástrojů je i vestavěný anotační nástroj. Anotace lze uložit do XML společně s anotovaným dokumentem, který je zde rozčleněn na identifikovatelné části (uzly). Každá anotace může mít seznam atributů, které mají název a hodnotu. Typy anotací i jejich atributy lze libovolně přidávat v průběhu anotování. Zobrazení je ve formě tabulky anotací pod dokumentem a blikání fragmentu textu patřícího k vybrané anotaci. Hlavními výhodami jsou rychlost a jednoduchost anotování, možnost dynamického přidávání typů i atributů a možnost libovolného překrývání fragmentů. Hlavní nevýhodou je, že se jedná o rozsáhlou desktopovou aplikaci, kterou je před anotováním potřeba spustit, importovat do ní dokument a následně jednotlivé typy anotací. Dalšími nevýhodami jsou nemožnost spolupráce v reálném čase a nemožnost tvorby odkazů mezi anotacemi.

Další editory

Existuje celá řada dalších řešení, která jsou určena převážně pro digitální knihovny (např. DiLAS [1]) nebo pro výzkumné účely. Příklady několika systémů lze nalézt v článkách Agosti, Ferra, da Rochy a kol. [1], [3], [21].

Kapitola 4

Zvolené metody řešení

V této kapitole jsou popsány zvolené metody řešení pro dosažení jednotlivých dílčích cílů práce.

Metodika srovnávání anotačních nástrojů

Při srovnání anotačních nástrojů nejprve začnu srovnáním většího množství nástrojů z obecnějších hledisek, jakými jsou typ aplikace, anotovaný objekt, typ obsahu anotace apod. Výslednou tabulku se srovnáním následně seřadím dle celkové vhodnosti jednotlivých nástrojů, přičemž bude brána v úvahu i důležitost jednotlivých srovnávacích kritérií pro následný postup. Několik vybraných nástrojů pak srovnám detailněji z hlediska pokročilé funkcionality (překrývání anotovaných fragmentů textu, typy atributů, vnořování anotací apod.). Dva nástroje, které ze srovnání vyjdou nejlépe, pak budou srovnány s nástrojem vytvořeným v rámci této práce při praktickém experimentu s vybranou skupinou uživatelů.

Při srovnávacím anotačním experimentu se vyhodnotí vhodnost jednotlivých nástrojů na komplexnější anotační úlohu a to, které aspekty uživatelského rozhraní jednotlivých nástrojů nejvíce ovlivňují rychlost anotování a kvalitu výsledků. Podstatné bude, do jaké míry a jakým způsobem je uživatel ovlivněn (zda např.

dochází k nezanedbatelnému nárůstu či poklesu jeho pozornosti apod.).

Anotační systém

Pro dosažení výše uvedených cílů bude třeba navrhnout nový anotační systém, se kterým bude možné provést experimenty.

Vývoj v oblasti anotací se dle Wenga a Gennariho [32] dělí na dvě hlavní kategorie. Jednou je vývoj datových schémat a druhou vývoj anotačních systémů. Vzhledem k výše uvedenému bude moje řešení v obou těchto oblastech, přičemž návrh formátu anotace spadá do první oblasti a návrh protokolu pro přenos anotací, klienta a serveru do druhé. Popsaný protokol pro přenos anotací jsem proto navrhl tak, aby jej bylo možné využít i s jiným formátem anotace.

Nový anotační systém bude tvořen serverem a klienty do různých prostředí. Při návrhu tohoto systému zvažím nejenom požadavky dle cílů práce, ale i další aspekty uvedené v literatuře, jako např. základní operace a procedury, které by měly anotační systémy poskytovat dle [4], a požadavky na anotace a anotační systémy dle [20]. Architektura systému bude navržena s ohledem na interoperabilitu a nahraditelnost jeho jednotlivých součástí.

Navržený protokol musí umožnit nejenom přenos samotných anotací oběma směry, ale také přenos typů anotací, změn v dokumentu a dalších důležitých informací pro pokročilé anotační úlohy. Obdobně jako popisují LaLiberte a Braverman [14] musí být možné jej bez další vrstvy přenášet v HTTP, což přináší určité problémy, protože protokol HTTP nebyl navržen pro obousměrnou asynchronní komunikaci.

Zvýšení kvality vytvářených anotací

Pro zvýšení kvality vytvářených anotací je potřebné zjistit, jak uživatelé při anotování postupují, na co se zaměřují a co naopak

snadno přehlednou. Je třeba vyhodnotit typy a příčiny chyb, které dělají, a navrhnout řešení, které umožní tyto chyby minimalizovat.

Za tímto účelem bude provedeno několik sad experimentů s uživateli. Při prvním experimentu budou provedeny testy s menším množstvím uživatelů, přičemž budu sledovat a vyhodnocovat jejich postupy. Po dokončení každého úkolu identifikuji chyby a s daným uživatelem prodiskutuji příčiny chyb a opatření, kterými by těmto chybám bylo možné předejít.

Následně budou připraveny experimenty se zaměřením na informace, které jsou uživateli zobrazeny, na způsoby jejich prezentace a na to, které informace zobrazit ihned a které až na vyžádání uživatelem.

Při anotování vždy dochází k určitému kompromisu mezi rychlostí anotování a kvalitou vytvářených anotací. Je tedy třeba se zabývat i usnadněním práce pro zvýšení rychlosti pomocí poloautomatického anotování a sémantického filtrování.

Kolaborativní tvorba ontologií

Při kolaborativním vytváření ontologií bude důležité navrhnout vhodné režimy interakce, umožňující takovou spolupráci znalce domény s odborníkem na vytváření ontologií, aby bylo dosaženo optimálního výsledku s minimálním úsilím.

Ve své práci se zaměřím na všechny tři fáze tohoto procesu. V první fázi vyhodnotím, do jaké míry mohou anotace v textu pomoci s tvorbou (počáteční verze) ontologie. Ve druhé fázi pak budu zkoumat, zda navržené řešení urychluje tvorbu informací pro vytvoření další verze ontologie. V závěrečné fázi pak vyhodnotím, zda kolaborativní anotační nástroj umožní urychlení a zefektivnění rozhovoru tvůrce ontologie se znalcem domény.

Kapitola 5

Realizovaný anotační systém

V první fázi práce byl navržen a implementován anotační systém, který byl nasazen v projektu Decipher. Po skončení projektu proběhla druhá iterace vývoje a s výsledným systémem byly prováděny další experimenty.

Anotační systém byl navržen a implementován s ohledem požadavky stanovené v úvodní části práce. Klienti jsou tvořeni doplňky do textových editorů a webových prohlížečů. Modulární server zajišťuje správu dokumentů, anotací a dalších potřebných informací. Pro komunikaci mezi anotačním serverem a klienty se využívá nově navržený protokol. Anotační server umožňuje export anotací do externích znalostních systémů v různých formátech.

Při zahájení práce v roce 2009 byly pro vytváření pokročilých anotací k dispozici pouze formáty Annotea [31] a Open Annotation [22], jehož vývoj byl právě zahájen. Jiné formáty využití v různých nástrojích nebyly navrženy univerzálně a nebyly ani dostatečně zdokumentované. Rozhodl jsem se proto, že navrhu vlastní formát, který bude později srovnán s právě vyvíjeným standardem Open Annotation.

V první verzi systému nasazené v projektu Decipher byl využit mnou navržený formát 4A. Ve druhé verzi systému byl následně nahrazen formátem Open Annotation. Oba formáty tak byly srovnány nejen teoreticky, ale i v praktickém nasazení.

Z hlediska jednoduchosti a rychlosti zpracování v anotačním systému je výrazně lepší formát 4A. U Open Annotation je třeba generovat řadu URI pro jednotlivé součásti anotace a zpracování musí mít více fází. Obtížnější jsou rovněž syntaktické kontroly, protože je zde velké množství elementů. S tím souvisí i vyšší efektivita přenosu dat při využití 4A. Z hlediska vyjadřovacích schopností poskytuje více možností formát Open Annotation. Kromě textu totiž umožňuje anotovat i multimediální data jako obrázky, videa apod. Nicméně v tomto ohledu by 4A bylo možné snadno rozšířit. Pro další použití výsledných anotací je rovněž výhodnější formát Open Annotation, neboť se jedná o čisté validní RDF, pro jehož zpracování existuje řada knihoven. Formát Open Annotation však neumožňuje vytvořit šablony pro anotace navázané na jednotlivé typy anotací. Vytvoření šablon tak vyžaduje netriviální externí rozšíření. V rámci testování jsem proto pro typy anotací využil specifikaci z formátu 4A.

Vývojem a následným testováním formátu 4A jsem prokázal, že lze vytvořit formát, který může úspěšně konkurovat Open Annotation a stále je zde tedy značný prostor pro zlepšení a vývoj nových formátů anotací. Ve výsledném systému vytvořeném v rámci této práce lze využít jak starší verzi klienta využívající 4A, tak i novou verzi využívající Open Annotation.

Pro přenos anotací mezi klientem a serverem jsem navrhl nový protokol, který kromě přenosu samotných anotací umožňuje přenos typů anotací a dalších potřebných informací pro pokročilé anotování. Umožňuje obousměrnou asynchronní komunikaci mezi klientem a serverem, takže pokud jeden uživatel zadá anotaci, server ji ihned pošle všem ostatním, kteří od daného uživatele odebírají anotace a anotují stejný dokument. Okamžitě se zasílají i aktu-

alizace typů anotací a obsahu dokumentu. Server může kdykoliv za běhu změnit i nastavení apod.

Oproti API, které navrhl Lee Feigenbaum ve firmě IBM [9], mnou navržený protokol předpokládá jednodušší formát anotace a zjednodušuje identifikaci v prostředí Internetu. Rovněž zaslání anotovaného textu na server a nabídnutí anotací je na rozdíl od tohoto API podporováno.

Pokročilé vlastnosti systému

Mezi klíčové cíle, kterými byl motivován vývoj systému, patří [26]:

- zvýšení úrovně detailu metainformací přesným ukotvením anotací v textu,
- umožnění práce v prostředí, na které jsou uživatelé zvyklí a využívají jej každý den,
- podpora objevování znalostí jako vedlejší efekt kolaborativního (sociálního) tagování,
- umožnění intuitivního strukturování znalostí ve formě přístupné běžnému uživateli,
- využití pokročilého zpracování textu a extrakce informací ke generování nabídek anotací a zvýšení efektivity procesu anotování.

Od výše uvedených cílů se odvíjí i základní pokročilé vlastnosti navrženého systému. K těm však přibyly i další klíčové vlastnosti, které rovněž vedou k usnadnění procesu anotování, zvýšení kvality vytvářených anotací a rozšíření možností nasazení systému:

- alternativní nabídky anotací,
- anotování při psaní textu,

- synchronizace pro kolaborativní anotování,
- pokročilé řízení odběrů anotací.

I přes to, že výsledné struktury anotací v RDF mohou být připojeny k celému dokumentu, je výhodné každou anotaci asociovat s konkrétním slovem, frází, větou či odstavcem, kterých se týká. Přesně ukotvená anotace je zásadní pro pozdější strojové zpracování anotovaných dat a využitelnost výsledných anotací. Metodám pro strojové učení umožňuje vytvářet lepší modely a při indexaci a následném vyhledávání umožňuje vytvořit náhled na správnou část dokumentu. Současně umožňuje i využití kombinace dotazování v jazyce SPARQL nad strukturami anotací v RDF s běžným plnotextovým vyhledáváním.

Systém má netriviální podporu překrývání fragmentů a to i částečného. Za účelem dodržení validity XML dokumentů a dalších hierarchických formátů dochází k automatickému dělení fragmentů na hranicích jednotlivých elementů dokumentu a k opětovnému spojování při prezentaci vytvořených anotací. Je tedy možné anotovat např. větu: „*Renoir and Manet made copies of Delacroix' paintings*“, kde identifikujeme dvě nezávislé události se dvěma různými osobami a sdíleným fragmentem reprezentujícím aktivitu (sloveso) a objektem vztahu.

K využití plného potenciálu navrženého systému je nezbytné, aby byli všichni zúčastnění motivováni aktivně přispět anotováním svého materiálu, zdrojových dat apod. Toho nelze dosáhnout, pokud anotování nebude extrémně jednoduché. Systém 4A proto vychází z myšlenky všudypřítomného anotování. Uživatelé nemusí měnit svoje zvyklosti, aby mohli anotovat v jednotném anotačním nástroji, ale anotování bude integrováno do nástrojů, které běžně využívají.

Sdílené koncepty ve formě ontologií jsou nutnou podmínkou k realizaci vize sémantického webu. I přes velkou snahu je však pokrytí vědeckých oblastí ontologiemi velmi omezené. Daný problém demonstruje nedostupnost široce akceptované ontologie v oblasti

počítačové vědy. Propojení procesu vytváření ontologie s procesem anotování skrývá tuto nepříjemnou práci a zdá se tak jako optimální strategie k překonání daného problému. Anotační systém 4A využívá konceptu sémantických šablon, které uživatele vedou skrze proces anotování. Efektivně se zde skrývá komplexita struktur ontologie před běžným uživatelem. Koncepty se zobrazují jako typy anotací využitelné v atributech a omezení na šablony atributů. Při jakémkoliv využití atributu je pak k dispozici odkaz, který jej váže zpět na původní ontologii. Toho pak lze využít i ke změnám a rozšiřování struktur dané ontologie.

Uživatelské rozhraní anotačního nástroje musí minimalizovat úsilí, které je potřebné k úspěšnému anotování obsahu. Pokročilé klasifikační mechanismy využitě pro nabízení anotací usnadní anotování základních kategorií a jejich zjednodušování s využitím kontextu a uživatel se tak může plně zaměřit na anotování specifické role, kterou entita v daném kontextu hraje.

Správnost nabídek anotací se výrazně liší dle obsahu daného dokumentu. Pro některé typy entit ve vhodném kontextu může být vysoká, pro jiné může být nízká a to zejména v případě komplexních a víceznačných popisů entit [17], [29]. Když nabídnutá anotace není korektní, uživatel ji může odmítnout a správnou anotaci vytvořit manuálně. Nicméně toto zahrnuje zdoluhavý úkol (vyhledání správné entity, ...). Abychom tento nepříjemný proces zjednodušili, do systému byla zabudována podpora pro alternativní nabídky anotací. Odmítne-li uživatel (nejlépe hodnocenou) nabídku pro daný fragment, systém mu nabídne sadu alternativ, ze které si může vybrat jedním kliknutím. Uživatel si rovněž může nechat zobrazit alternativy bez odmítnutí nejlépe hodnocené nabídky, aby se mohl ujistit o správnosti jejího potvrzení.

Když uživatel napíše novou část textu v textovém editoru, může tento text ihned anotovat. Text a anotace jsou zpracovány odděleně, což umožňuje lepší podporu současného editování a anotování textu. Pokročilé metody správy anotací pak garantují, že většina anotací zůstane po editaci stále validních a pouze skutečně

vyřazené anotace budou zneplatněny. Anotace jsou v průběhu editace průběžně aktualizovány anotačním serverem, který k tomuto účelu využívá kaskádu metod s různou citlivostí. V průběhu psaní dochází i k aktualizaci nabídek anotací a než tak uživatel dopíše větu, jsou mu k ní již nabídnuty anotace zmíněných entit a dalších rozpoznatelných prvků.

Abychom se vyhnuli problémům při paralelní práci uživatelů a současně aby mohli lépe spolupracovat, systém má podporu pro synchronizaci změn v dokumentu, anotacích i znalostních strukturách v reálném čase.

Pokročilé řízení odběrů anotací umožňuje definovat sady pravidel stanovujících zdroje a typy odebíraných anotací. Tyto sady nazýváme odběry a lze je libovolně kombinovat. Toto lze provést separátně pro každou instanci editoru (např. v různých záložkách prohlížeče) a získat tak různé pohledy na anotovaný text, které usnadní vyhodnocení odpovídajících aspektů řešeného problému.

Kapitola 6

Provedené experimenty

Všechny experimenty byly zaměřeny především na pokročilé anotační úlohy, jako je hierarchické anotování komplexních vztahů v textu a entit s mnohoznačnými jmény a tomu odpovídajícími mnoha potenciálními položkami ve znalostní bázi. Hlavním účelem experimentů bylo sledování, jak určité aspekty uživatelského rozhraní anotačních nástrojů ovlivňují rychlost a kvalitu poloautomatického procesu anotování. Výsledky experimentů byly publikovány v několika článcích [26], [27], [7], [8].

Nejprve byly provedeny úvodní experimenty, jejichž cílem bylo vyzkoušet první verzi prototypu realizovaného anotačního nástroje a ověřit základní koncepty, jako je způsob výběru anotovaných fragmentů a kolaborativní vytváření znalostních struktur. Výsledky byly publikovány v článku *Towards New Scholarly Communication: A Case Study of the 4A Framework* [26].

Následovalo praktické nasazení v projektu Decipher, z nějž máme vyhodnocení z perspektivy uživatelů a jejich pohledu na jednotlivé prvky uživatelského rozhraní, které se ukázalo jako klíčové pro vývoj druhé verze systému.

První série experimentů po skončení projektu Decipher byla zaměřena na následující zkoumané otázky:

1. Jak jednotlivé prvky uživatelského rozhraní ovlivňují kvalitu výsledků a rychlost anotování?
2. V jakém rozsahu množství zobrazovaných informací ovlivňuje výsledky zjednodzračňování entit?
3. Je pro uživatele výhodou znalost potenciálních alternativních anotací a poskytnutí míry důvěryhodnosti jednotlivých nabídek anotací?
4. Jaké výhody do procesu anotování přináší koncept sémantického filtrování?

Vysoký počet existujících anotačních systémů je v přímém kontrastu s faktem, že existuje pouze velmi málo studií, které se zabývají jejich srovnáním z hlediska funkcionality a využitelnosti pro konkrétní úlohy. Úvodní experiment z této série byl zaměřen na první uvedenou otázku. Bylo provedeno srovnání různých uživatelských rozhraní a vzorů interakce reprezentovaných třemi různými anotačními systémy – 4A vyvinutým v rámci této práce, RDFaCE [13] a Gate Developer [6] – na úloze hierarchického anotování komplexních vztahů v textu. Tyto nástroje byly zvoleny dle výsledků předchozího teoretického srovnání většího množství anotačních nástrojů. Proces anotování se skládal z volby části textu korespondující s událostí specifického typu a vyplnění atributů entitami a hodnotami zmíněnými v textu. Entity je při tom nutné zjednodzračnit jejich spojením s referenčními zdroji (převážně Wikipedia a DBPedia).

Pro minimalizaci rizik nespravedlivého porovnání bylo pro experimenty zaměřené na srovnání anotačních nástrojů zvoleno 6 uživatelů, kteří neměli žádné předchozí zkušenosti s testovanými nástroji, s řešenými úlohami, ani žádné zvláštní znalosti z domény anotovaných textů. Každý strávil 20 minut seznámením s daným nástrojem, přičemž pro zvýšení spravedlivosti srovnání se pro každého uživatele lišilo pořadí, ve kterém byly nástroje testovány. V průběhu experimentu měl každý uživatel 40 minut na anotování

každým z nástrojů. Instrukce požadovaly, aby se uživatelé snažili dosáhnout maximální možné přesnosti, ale čas na dokončení celého úkolu byl omezený. Byly při tom měřeny tři hodnoty: Jak je uvedeno v tabulce 6.1, patří mezi ně množství nesprávně vyplněných hodnot atributů, počet chybějících atributů a průměrný čas na anotaci události. Nesprávně vyplněné atributy při tom zahrnovaly všechny druhy chyb (např. nesprávný typ, koreference, URL apod.).

Tabulka 6.1: Výsledky experimentu zaměřeného na srovnání anotčních nástrojů

Nástroj	Nesprávné hodnoty	chybějící hodnoty	čas na událost
GATE	9.4 %	8.3 %	135 s
RDFaCE	8.7 %	8.8 %	193 s
4A	6.2 %	5.6 %	116 s

Celkový vysoký počet chyb lze vysvětlit velmi striktním porovnáním s referenčními anotacemi. Výsledky korespondují s tím, že způsob, jakým GATE Developer prezentuje anotace atributů událostí, často vede k nejasným či nekonzistentním výsledkům. Atributy jsou totiž k události přiřazeny pouze faktem, že fragment označující událost překrývá fragment s hodnotou atributu. RDFaCE je v tomto ohledu jen o málo lepší, protože sice umožňuje přiřadit atribut k dané události, ale u překrývajících se událostí problémy přetrvávají.

Ve výše uvedených výsledcích se jasně projevil rozdíl mezi RDFaCE či GATE a 4A, který ukazuje, že je přínosné vizuální odlišení nabídek anotací generovaných systémem od anotací zkontrolovaných či vytvořených uživatelem. Způsob potvrzování v systému 4A tedy vede na lepší a konzistentnější anotace. Rozdíl časů

lze mimo jiné vysvětlit i tím, že u 4A je využito sémantické filtrování, které urychlí práci uživatele.

Další dvě z výše uvedených otázek jsou pokryty druhou sadou experimentů. Je jasné, že množství zobrazovaných informací a způsob jejich zobrazení mohou ovlivnit rychlost a přesnost anotování. Experimenty využívají nástroj 4A ke studii srovnávající různá nastavení uživatelského rozhraní anotačního nástroje a souvisejících vzorů interakce uživatelů. Ukazuje se, že praxe běžně využívaná v anotačních nástrojích, která požaduje zjednodušování entit pouze na základě typu a zobrazovaného URL, vede k nízké kvalitě výsledků. Opačný přístup, kdy uživatelé okamžitě dostanou rozsáhlé informace o daných entitách potenciálně korespondujících s daným víceznačným jménem, však vede nejen k delšímu času anotování, ale překvapivě ani nezvyšuje kvalitu vytvářených anotací. Je tedy potřeba najít určitý kompromis mezi nedostatečnými a příliš rozsáhlými informacemi. Takové optimální nastavení parametrů bylo v rámci experimentů vyhledáváno.

Primárně byla srovnána 3 nastavení pohledu pro zjednodušování, která se lišila zobrazenými atributy, a byl sledován dopad daného nastavení na rychlost a přesnost zjednodušování. Uživatelé dostali instrukce, aby anotovali pouze zadanou entitu, která byla v každém úryvku textu zvýrazněna, a zvolili vždy právě jednu z poskytnutých nabídek anotací. V prvním nastavení měli uživatelé pro každou nabídku s nejvyšší mírou důvěry zobrazený rozsáhlý seznam atributů a jejich hodnot. Druhé nastavení korespondovalo s omezeným zobrazením, které pro zjednodušování poskytuje celá řada nástrojů (např. RDFaCE). Byl tedy zobrazen pouze typ entity a URL. Uživatel se pak mohl rozhodnout podle samotné URL (na Wikipedii často obsahuje slovo či frázi usnadňující zjednodušování), nebo si zobrazit cíl daného odkazu. Třetí nastavení umožňuje využít výhodu speciálního zjednodušovacího atributu, který je automaticky určen z popisů dostupných alternativ. Společně s tímto atributem byl zobrazen i typ entity a URL.

Zatímco sekvence testovacích textů (40 pro každé nastavení) zůstala fixní, každý ze šesti testerů měl jiné pořadí tří uvedených nastavení. Každý měl 30 minut na každé nastavení. V tabulce 6.2 je srovnání časů, chybovosti a toho, jak moc kliknutí na URL a prohlédnutí příslušné stránky s dalšími informacemi v referenčním zdroji uživatelé potřebovali.

Tabulka 6.2: Výsledky experimentu srovnávající tři různá nastavení zobrazení pro zjednodušování

nastavení	průměrný čas	množství chyb	kliknutí na URL
detailní informace	33.92 s	6.2 %	1.3 %
pouze typ a URL	37.26 s	27.9 %	41.7 %
zjednodušovací atribut	32.98 s	2.1 %	1.5 %

I když mezi jednotlivými testery byly určité rozdíly, celkové výsledky byly pro všechny testery stejné. Počet případů, ve kterých uživatel potřeboval informace z kompletní stránky v referenčním zdroji, byl vždy nejvyšší při druhém nastavení (pouze typ a URL). Jako jasně nejlepší nastavení v této části experimentů se ukázalo využití zjednodušovacího atributu s možností kliknutí na URL pro zobrazení detailů. Uživatelé s tímto nastavením udělali méně chyb než s ostatními a dosáhli rovněž nejnižšího průměrného času na anotaci.

Bez podpory systému se uživatel nemusí dozvědět, že je dané jméno víceznačné. Pokud uživatel není expertem na danou oblast a jsou zde dva nebo více potenciálních významů, může jednoduše potvrdit nabídnutý nesprávný význam. Toto riziko lze snížit tak, že nástroj uživatele upozorní na existenci alternativ. Otázkou je, jak má vypadat optimální konfigurace této funkcionality. Tedy kdy se mají alternativy zobrazit. Pro experiment zkouma-

jící tento problém byly připraveny dvě datové sady po 15 úryvcích textu zmiňujících vybrané entity, kde 3 z nich (20 %) v každé sadě nebyly pokryty referenčními zdroji a tedy nekořespondovaly s žádnou z nabídnutých alternativ. První nabídka byla správná v 9 z 15 případů v každé sadě (60 %). První testované nastavení kořespondovalo s obecně využívanou prezentací nejlepší alternativy anotace se zobrazením zjednoznačňovacího atributu. Druhé nastavení ihned zobrazilo 5 nejpravděpodobnějších nabídek a míru důvěry ve správnost nejlepší z nich. Tři uživatelé začali s prvním uvedeným nastavením, 3 s druhým. V případech, kdy nebyla žádná z nabídek správná, se za správné rozhodnutí považovalo odmítnutí všech nabídek. Výsledky jsou shrnuty v tabulce 6.3.

Tabulka 6.3: Srovnání zobrazení jedné nabídky anotace a všech alternativ

nastavení	průměrný čas	množství chyb	kliknutí na URL
jedna nabídka	41.2 s	14.4 %	3.3 %
všechny alternativy	42.9 s	12.2 %	2.2 %

Přesto že výsledky nejsou zcela přesvědčivé a obecně závisí na kvalitě procesu generování nabídek anotací v pozadí a na četnosti případů, kdy zmíněná entita není pokryta referenčními zdroji, zobrazení všech alternativ se ukazuje jako vhodnější, chceme-li v daném čase preferovat kvalitu vytvářených anotací před kvantitou.

Poslední sada experimentů v této sérii následně prokazuje, že sémantické filtrování implementované v nástroji 4A zrychluje proces anotování událostí a přináší vyšší konzistenci než alternativní přístupy. Byly připraveny dvě sady textů, z nichž každá obsahovala 6 událostí správného typu. V první sadě textů měli uživatelé identifikovat atributy událostí manuálně, ve druhé sadě bylo využito automatické předanotování potenciálních hodnot atributů všech

potřebných typů (nabídky anotací) a pokročilé sémantické filtrování systému 4A, které zvýraznilo potenciální kandidáty na hodnoty právě vyplňovaného atributu.

Výsledky pro jednotlivá nastavení jsou porovnány v tabulce 6.4. Manuální proces bez předanotování je zdoluhavý a uživatelé tak strávili více než 5 minut anotováním jedné události. Výsledky také obsahovaly mnoho chyb. Oproti tomu nabízení anotací a sémantické filtrování zapnuté ve druhé části experimentu vedly k dosažení kvalitních výsledků anotování v relativně krátkém čase.

Tabulka 6.4: Manuální vyhledávání hodnot vs. nabízení anotací se sémantickým filtrováním

vyplněné atributy	nesprávné hodnoty	chybějící hodnoty	čas na událost
manuálně	11.7 %	6.6 %	303.5 s
s nabízením	4.5 %	3.4 %	109.3 s

V další sérii experimentů pak bylo blíže zkoumáno sémantické filtrování a optimalizace množství zobrazovaných informací a byly zkoumány přínosy alternativních nabídek anotací, aby bylo možné zodpovědět některé otevřené otázky z předchozích experimentů. Oproti předchozí sérii experimentů byl zvýšen počet anotátorů a pro získání nových pohledů na vzory interakce byly připraveny nové anotační úlohy.

Výsledky předchozích experimentů ukázaly, že i přes to, že automatický proces anotování nemůže identifikovat komplexní vztahy, je výhodné jej využít pro předanotování entit a základních vztahů. Z výsledků však nebylo zcela jasné, v jakém rozsahu zvýraznění preferovaných typů anotací v textu využitelných jako hodnoty atributů napomáhá k vyšší kvalitě a rychlosti anotování vztahů a zda zvyšuje komfort uživatelů. Pro experiment, který měl tento rozsah určit, byla připravena sada 20 úryvků z dokumentů o vizuálních

uměleckých dílech a vztazích vzájemného ovlivnění umělců, včetně ovlivnění mezi těmito díly. Skupina 14 uživatelů v textech nejprve identifikovala umělecká díla a jejich atributy a následně anotovali vztahy ovlivnění mezi těmito díly a jejich autory. Byly připraveny 2 různé konfigurace systému, z nichž s každou uživatelé anotovali 10 textů. Při první konfiguraci byly zvýrazněny anotace příslušných typů korespondující s možnými hodnotami právě vyplňovaného atributu. Při druhé konfiguraci bylo sémantické filtrování vypnuté.

V tabulce 6.5 je uvedeno srovnání výsledků získaných s jednotlivými variantami nastavení. Je zřejmé, že zapnuté sémantické filtrování nástroje 4A vedlo k vyšší kvalitě výsledků. Relativní snížení obou typů chyb překročilo 25 %. Proces anotování byl také o 15 % rychlejší. Dotazníky vyplňované po skončení experimentu pak ukázaly, že 11 ze 14 uživatelů je toho názoru, že funkce sémantického filtrování výrazně zlepšuje jejich komfort při anotování.

Tabulka 6.5: Přínos sémantického filtrování

sémantické filtrování	nesprávné hodnoty	chybějící hodnoty	čas na vztah
vypnuté	6.9 %	5.7 %	41.4 s
zapnuté	5.1 %	4.2 %	35.1 s

V dalším experimentu, rozšiřujícím předchozí experiment zaměřený na alternativní nabídky anotací, byly srovnány dvě možnosti prezentace seznamu alternativ uživateli. Ve stručnější variantě byly vypsané pouze zjednoznačňovací kontexty a umožněno rozbalení kompletního seznamu atributů entity, ve druhé variantě pak byly všechny atributy ihned rozbalené a uživatel měl možnost je pro každou alternativní nabídku jedním kliknutím sbalit. Každý ze 20 zjednoznačňovacích úkolů byl řešen 7 anotátory se stručným pohledem a 7 s rozbaleným pohledem.

Souhrnné výsledky tohoto experimentu jsou uvedeny v tabulce 6.6. Stručný pohled uživatelům umožnil rychleji procházet alternativy a zvolit jednu z nich. Všichni kromě jediného uživatele zde jako preferovaný označili právě tento pohled.

Tabulka 6.6: Zjednodušování se stručným a s rozbaleným pohledem na alternativní nabídky anotací

nastavení	čas na alternativu	množství chyb	kliknutí na sbalení /rozbalení
sbalené	11.2 s	12.1 %	1.1 %
rozbalené	15.1 s	11.4 %	5.3 %

Poslední experiment, který bude popsán v této kapitole, se zabýval kolaborativním vytvářením ontologie pro aspektově orientovanou analýzu sentimentu. Pro rozšiřování ontologie byly využity dva různé nástroje – 4A a WebProtégé¹. Vždy, když uživatelé vytvořili nový koncept v nástroji 4A, ihned pomocí něj anotovali zmínky o jednotlivých aspektech, které se vyskytovaly ve zpracovávaných textech recenzí. Při využití WebProtégé však texty sloužily pouze jako doplňkový zdroj znalostí. Proběhly 4 běhy kolaborativní přípravy ontologie. Dvě dvojice (první znalostní inženýr s 1. informovaným klientem v prvních 2 bězích a s druhým ve druhých 2 bězích) vždy dostaly stejnou sadu textů ale pracovaly buď s 4A, nebo s WebProtégé. Dokumenty s recenzemi byly nejprve dány znalostním inženýrům, kteří si je přečetli, aby se seznámili s jednotlivými aspekty názorů zmíněných v reálných datech. Měřený úkol se pak v každém běhu skládal ze 2 kroků. V prvním kroku byla zahrnuta individuální práce obou účastníků. Druhý krok zahrnoval společnou práci celé dvojice. V prvním kroku znalostní inženýr připravil svůj první výchozí návrh ontologie s využitím

¹<http://webprotege.stanford.edu/>

přiřazených nástrojů, zatímco informovaný uživatel dostal za úkol pokusit se vypsát potenciální aspekty v dané oblasti v prostém textu. Ve druhém kroku každá dvojice porovnála výchozí návrh ontologie s neformálním seznamem aspektů, identifikovala konfliktní pohledy a výchozí ontologii modifikovala tak, aby vyjadřovala od-souhlasené společné porozumění dané oblasti.

V tabulce 6.7 jsou shrnuty průměrné časy k dokončení popsa-ných dvou kroků s využitím obou nástrojů. I když se čtyři výsledné ontologie výrazně liší v počtu aspektů, na kterých se uživatelé do-hodli, aby mohly být zahrnuty ve výsledné struktuře, obecný vzor ukázaný v tabulce zůstal přes jednotlivé běhy stejný. Přidaná práce spojená s anotací příkladů v textu prokládaná doplňováním struk-tury znalostí v nástroji 4A vedla k mírnému zpomalení individu-ální přípravy ontologie znalostními inženýry. Dostupnost reálných dat anotovaných navrženou ontologií však výrazně zrychlila řešení identifikovaných konfliktů a redukovala potřebu vysvětlovat a dis-kutovat kategorizaci každého jednotlivého aspektu.

Tabulka 6.7: Výsledky experimentu na srovnání nástrojů pro roz-šířování ontologií

nástroj	průměrný čas	
	individuálního kroku	kolaborativního kroku
WebProtégé	18.6 min.	15.3 min.
4A	19.1 min.	12.6 min.

Kapitola 7

Dosažené výsledky a jejich analýza

V této kapitole shrnu nejdůležitější dosažené výsledky této práce a analyzuji jejich potenciální dopad na aktuální stav poznání a budoucí vývoj v dané oblasti.

Mezi hlavní technické výsledky patří navržený protokol pro přenos anotací a realizovaný anotační systém, který přináší nejen řadu nových konceptů a inovací oproti existujícím systémům, ale poskytuje také platformu pro testování různých aspektů uživatelských rozhraní anotačních nástrojů.

Následné experimenty pak prokázaly správnost konceptů optimalizujících množství zobrazovaných informací, čímž lze dosáhnout vyšší rychlosti anotování i kvality vytvářených anotací.

Zvláštní důraz byl kladen na prozkoumání nového konceptu alternativních nabídek anotací a konceptu sémantického filtrování. V závěru jsem pak zkoumal výhody realizovaného anotačního nástroje pro kolaborativní vytváření ontologie.

7.1 Protokol pro přenos anotací a formát anotace

Dosud neexistoval žádný otevřený protokol pro přenos anotací, který by umožňoval komplexnější komunikaci než pouhé uložení, aktualizaci či smazání samotné anotace. API, které navrhl Lee Feigenbaum ve firmě IBM [9], má oproti protokolu 4A značnou nevýhodu v tom, že je potřeba knihovna, která na straně klienta zajistí komunikaci se serverem skrze dané API. Při definování jasného protokolu pak lze jednoduše implementovat klienta, který nebude závislý na konkrétní knihovně, ani na platformách, pro které byla knihovna implementována.

Použitelnost a univerzálnost navrženého protokolu pro přenos anotací byla prokázána při nasazení v evropském projektu Decipher. Změna ze mnou navrženého formátu na formát Open Annotation přitom nevyžadovala žádný zásah do základu protokolu – změny ve verzi 2.0 oproti 1.0 byly motivovány pouze nedostatkami ve verzi 1.0 zjištěnými při nasazení v projektu Decipher a požadavky na nově přidávanou funkcionalitu. Tím byla potvrzena možnost univerzálního využití protokolu pro různé formáty anotací.

Navržený formát anotace 4A se při testování ukázal jako alternativa k formátu Open Annotation umožňující jednodušší a efektivnější komunikaci mezi klientem a serverem i jednodušší zpracování dat. I přes to, že formát Open Annotation ve své nejnovější verzi přejmenované na Web Annotation [23] poskytuje více možností než mnou navržený formát a jednodušší zpracování výsledných anotací různými nástroji, protože se jedná o standardní RDF, stále nenabízí vhodnou specifikaci pro popis šablon pro atributy jednotlivých sémantických typů anotací. Tím je prokázáno, že je zde stále prostor pro vylepšení i pro vývoj nových formátů anotací umožňujících jednodušší zpracování, které by bylo výhodou pro mobilní platformy. Do budoucna by na základě formátu 4A mohlo být vytvořeno rozšíření formátu Web Annotation, které by rozší-

řilo jeho možnosti pro využití pro komplexní anotační úlohy (např. anotování vztahů).

Do budoucna mám v plánu navržený protokol a rozšíření formátu Web Annotation nabídnout kolegům ze skupiny *W3C Web Annotation Working Group*, aby mohl být masivněji rozšířen a využíván. Kdyby v budoucnu došlo ke standardizaci protokolu pro přenos anotací, umožnilo by to širokou interoperabilitu anotačních nástrojů s modelem klient – server a tím i rychlejší vývoj v oblasti sémantického anotování textu.

7.2 Vytvořený anotační systém

Na vytvořený anotační systém se lze dívat z více úhlů pohledu. Z praktického pohledu jej lze chápat jako univerzální anotační nástroj, který lze snadno nasadit do systémů pro správu obsahu a využít pro hierarchickou anotaci komplexních událostí. Použitelnost k tomuto účelu byla prokázána v projektu Decipher.

Hlavním cílem nástroje však bylo připravit platformu umožňující výzkum v oblasti anotování. Celá řada konceptů byla v minulosti otestována na nástrojích, které byly navrženy a implementovány primárně pro otestování daného konceptu. Např. Zheng navrhl Bundle Editor [34] pro ověření konceptu strukturování anotací a Yee a jeho tým implementovali CritLink [33], aby získali nástroj s kombinací všech funkcí, které identifikovali jako základ pro úspěšný editor anotací. Znovupoužitelnost takových nástrojů pro ověření dalších konceptů však byla často malá a celá řada nástrojů posloužila pouze k jednorázovému výzkumu, po kterém byla jejich podpora ukončena. Některé nástroje pak skončily v počátcích pokusu o komerční provoz (např. SharedCopy [25]). Implementace základu anotačního nástroje je však náročná, přičemž se složitostí struktur anotací a jejich vizualizace značně narůstá. I pro ověření poměrně jednoduchých konceptů je tak vynakládáno značné úsilí na vývoj nástrojů, ve kterých tyto koncepty bude možné ověřit.

Nástroj 4A byl proto vyvinut tak, aby byl v maximální možné míře modulární, konfigurovatelný a opakovaně využitelný pro ověřování nových konceptů. Např. při přidání alternativních nabídek anotací tak postačovalo přidat jeden modul serveru a mírně rozšířit komunikační protokol, modul pro komunikaci a klienta, aniž by byly potřebné výraznější zásahy do základní funkcionality systému. Vyřazení alternativ pak lze provést pouhým vypnutím či vyřazením příslušného modulu na serveru. Pro níže uvedenou optimalizaci množství zobrazovaných informací dokonce nebylo potřeba provádět žádné změny implementace, neboť vše potřebné bylo možné řešit úpravami konfigurace. Vzhledem k tomu, že je systém navržen i pro práci s komplexními hierarchickými anotacemi, nabízí široké spektrum možností anotování a poskytuje tak vynikající základ pro rozšiřování a testování dalších konceptů. Výsledné řešení je licencováno svobodnou licencí (Apache¹) a může být dále využíváno výzkumníky, kteří budou objevovat nové koncepty a budou potřebovat systém, ve kterém by je mohli otestovat a vyhodnotit.

Využitelnost systému 4A pro ověřování nových anotačních konceptů byla demonstrována provedenými experimenty, jejichž vyhodnocení následuje níže. Vytvořený systém mi tak umožnil dosažení dalších cílů mé práce.

7.3 Vyhodnocení srovnání anotačních nástrojů

Dle kritérií uvedených v kapitole 4 jsem provedl srovnání anotačních nástrojů. Na toto teoreticky zaměřené srovnání navazoval praktický experiment srovnávající vybrané pokročilé anotační nástroje popsány v kapitole 6.

Z tabulek v příloze mé práce je patrné, že existuje velké množství nástrojů. Některé nástroje již nejsou dostupné a některé nové

¹<http://www.apache.org/licenses/LICENSE-2.0.txt>

do tabulky dosud nebyly zařazeny. Je však překvapivé, že novější nástroje často nepřinášejí rozšířenější funkcionalitu, ale pouze jiné způsoby, jak realizovat základní funkce. Jen omezené množství nástrojů nabízí možnost strukturování anotací a pouze málo z nich umožňuje strukturování jiným způsobem než pouze ve využití ontologii, kterou drtivá většina nástrojů neumožňuje přímo modifikovat.

Ze zastoupení webových aplikací a doplňků do webových prohlížečů je patrný trend vedoucí k anotování v prostředí webu a aplikací, které uživatelé běžně využívají. To potvrzuje správnost myšlenky, na které je postaven nástroj 4A.

Poněkud neutěšenou situaci lze sledovat v oblasti využitých formátů anotací a protokolů využívaných pro jejich přenos. Standardizované formáty jako Open Annotation, Annotea či RDFa (který však vyžaduje přenos anotovaného textu s anotacemi) zatím využívá jen malé množství nástrojů. Další nástroje pak využívají standardní obecné formáty jako XML či JSON, ale struktura dat je zde určena autorem systému a nejsou tak vzájemně interoperabilní. Téměř pětina nástrojů pak pracuje s uzavřeným proprietárním formátem, což jakoukoliv otevřenou interoperabilitu zcela vylučuje. Za dobu svého výzkumu však pozoruji jasný trend vedoucí k širšímu využití formátu Web Annotation, který je patrný i z vícenásobného zastoupení jeho předchůdce Open Annotation v řadě nástrojů.

Pro přenos anotací je nejvíce využíván protokol HTTP. Tím jsou pak přenášeny přímo anotace (často se využívá jednoduché REST API²), případně jednoduchý proprietární protokol na nižší úrovni. Definovaný protokol na vyšší úrovni pak využívá jen zanedbatelné množství nástrojů, které definují několik typů zpráv ve formátu JSON. Interoperabilita nástrojů je díky nejednotnosti protokolu značně komplikovaná. Zde se nabízí prostor pro mnou navržený protokol 4A.

²Representational state transfer <http://www.restapitutorial.com/>

Srovnání pokročilých nástrojů rovněž ukazuje, že i nejpokročilejší z dostupných nástrojů mají značné rezervy. Některé z nich neumožňují vytvoření odkazů mezi anotacemi, které je potřebné pro jejich strukturování, a některé tuto činnost nemají intuitivní pro laiky (přímé skládání trojic RDF). Vnořování pak většinou není umožněno, nebo je pouze omezené. Problémem je i překrývání anotací, které je potřebné při anotování dvou nezávislých událostí popsaných v jedné větě (např. když různí malíři zcela nezávisle vytvořili obrazy na dané téma nebo dokonce pouze stejnou technikou). Nabízení anotací s odlišením nabídek od uživatelských anotací kromě mnou navrženého nástroje z pokročilých nástrojů umí pouze nástroj Domeo [5], přičemž vedoucí týmu, který vytvořil tento nástroj (Paolo Ciccarese), je současně jedním z klíčových expertů ve skupině, která navrhla formát Web Annotation. Tyto výsledky poukazují na fakt, že je dostupných jen velmi málo pokročilých univerzálních anotačních nástrojů a jejich funkcionalita je často omezená, což je dané primárními případy použití, pro které byly navrženy. Je zde tedy stále velký prostor pro vylepšení, do kterého zasahuje i mnou navržený nástroj 4A.

Praktický experiment srovnávající tři vybrané nástroje pak zkoumal, zda a jakým způsobem jednotlivé aspekty uživatelského rozhraní jednotlivých anotačních nástrojů ovlivňují práci uživatelů. Výsledky ukazují, že příliš strohé uživatelské rozhraní, které nenabízí intuitivní provádění základních operací s podporou systému, vede na vysoké procento chyb ve vytvářených anotacích i na degradaci rychlosti anotování. Rovněž způsob prezentace příslušnosti atributů ke komplexní anotaci pouhým překrytím fragmentů vede k nejistotě uživatele a zvýšení počtu chyb. Naopak přehledné uživatelské rozhraní odlišující automaticky vygenerované nabídky anotací od uživatelských anotací a vhodný způsob prezentace atributů komplexních vztahů umožňují zvýšit rychlost anotování i kvalitu vytvářených anotací. Vzhledem k tomu, že při anotování uživatel volí kompromis mezi rychlostí práce a kvalitou vytvářených anotací, je zlepšení v obou těchto kritériích, kte-

rého nástroj 4A dosáhl díky zmíněným rozdílům, nezanedbatelné, i když je v řádu jednotek procent a jednotek až desítek sekund na komplexní anotaci události. V každém případě však prezentované výsledky jasně prokazují, že uživatelské rozhraní zásadně ovlivňuje nejen rychlost anotování, ale i kvalitu vytvářených anotací, což bylo jedním z cílů mojí práce.

Srovnání nástroje umožňujícího strukturování pouhým překrytím fragmentů s nástrojem využívajícím strukturované anotace pak prokázalo, že strukturované anotace umožní lepší popis sémantiky a poskytují možnosti lepšího strojového zpracování vhodného mimo jiné i pro jejich následnou vizualizaci.

7.4 Určení optimálního množství zobrazených informací

Ve výše popsaném experimentu byla srovnána 3 různá nastavení zobrazení atributů nabídek anotací v anotačním nástroji 4A a sledován dopad jednotlivých nastavení na rychlost a přesnost zjednoznačňování entit. Výsledky prokázaly, že často využívaný typ zobrazení obsahující pouze typ a URL entity v referenčním zdroji vede na největší množství chyb ve vytvořených anotacích. Vzhledem k časté potřebě prohlížení obsahu stránky, na kterou URL odkazuje, současně dochází ke zpomalení procesu anotování.

Následně však bylo prokázáno, že zobrazení většího množství informací, než je nezbytně nutné, vede nejen ke zpomalení procesu anotování, ale i k nárůstu počtu chyb. Uživatelé sice mohou s výhodou hledat klíčová slova z kontextu dané zmínky o entitě v popisu nabízených alternativ, ale velké množství zobrazených informací rovněž rozptyluje jejich pozornost a často se nesprávně rozhodnou na základě nepřesné či nerelevantní informace (např. informace, která je pro dvě zjednoznačňované entity společná, ale uživatel se již při nalezení prvního výskytu nesprávně rozhodne, protože se mu daná informace zdá dostatečně specifická).

Nalezení optimálního množství zobrazovaných informací je obtížnou úlohou a je jen velmi těžko generalizovatelné, neboť se liší pro jednotlivé oblasti zájmu (historie, biomedicína, chemie apod.). Provedené experimenty byly proto zaměřené na zjednodušování mnohoznačných názvů osob a míst, protože je tento typ zjednodušování velmi častým a uplatňuje se mimo jiné i v doménách (nejen evropských) projektů, kterými se zabývá Výzkumná skupina znalostních technologií na FIT VUT v Brně. Při této úloze je k dispozici celá řada informací od data narození konkrétní osoby, bydliště a dalších osobních informací, přes popis její činnosti a zahraničních cest až po výsledky její práce (např. umělecká díla). Pro člověka je volba nejdůležitějších informací intuitivní, zatímco pro stroj je obtížné určit optimální algoritmus, který dosáhne uspokojivých výsledků. Ideální zjednodušovací informace by obsahovala pouze to, co je podstatné pro rozhodnutí – tedy to, v čem se jednotlivé alternativy liší a je to přímo uvedeno v textu. Kdybychom takovou informaci měli, byla by úloha značně zjednodušená a pravděpodobně by ji bylo možné provádět i automaticky. V praktických příkladech je však situace složitější, neboť v dostupných popisech entit, které jsou převážně v přirozeném jazyce, není jednoduché identifikovat klíčové aspekty pro vyhodnocení rozdílů. Potřebná informace z textu často pouze plyne, ale není v něm přímo uvedena (např. je v textu zmínka o tom, že osoba nebude kandidovat na prezidenta, a z popisu entity víme, že se jedná o politika, aniž by slovo prezident či politik bylo přímo zmíněno, zatímco ostatní alternativy politikem nejsou). V takovém případě je tedy třeba zvolit určitý kompromis, kdy jsou odhadnuty nejdůležitější informace a jsou-li nedostačující, uživatel musí využít celý popis entity a dohledat chybějící fakta. Námi testovaný zjednodušovací atribut byl proto vytvořen ze zjednodušovacích informací v URL na anglickou wikipedii, které jsou vytvářeny lidmi, a vybrané části popisu entity z Wikipedie, kde se autoři obvykle snaží o jasný a výstižný popis. Obdobně lze postupovat i pro jiné oblasti – najít vhodný referenční zdroj a určit algoritmus výběru zjedno-

značňovací informace. Námi vyvíjený komponent pro sémantické obohacení textu má architekturu umožňující snadné přizpůsobení pro libovolné textové referenční zdroje.

Nastavení s navrženým zjednoznačňovacím atributem v experimentu i přes relativně jednoduchý algoritmus generování zjednoznačňovací informace vedlo k dosažení nejlepších výsledků ze všech srovnávaných nastavení a to nejen co se týče kvality vytvořených anotací, ale i rychlosti anotování. Právě zobrazením zjednoznačňovacího atributu a možností rozbalení dalších detailnějších informací, ze kterých se pomocí URL lze dostat až k plnému rozsahu informací dostupných v referenčním zdroji, se mi tedy v anotačním nástroji 4A v rámci možností podařilo optimalizovat množství zobrazovaných informací tak, že jsem zvýšil rychlost anotování i kvalitu vytvářených anotací oproti přístupům využívaných v nejlepších dostupných existujících nástrojích, se kterými byl nástroj 4A srovnán ve výše popsaném experimentu.

7.5 Přínos alternativních nabídek anotací

Srovnávané nástroje, které umožňují poloautomatické anotování textu, při předanotování textu či vygenerování nabídek anotací vytvoří pro každý fragment textu zmiňující nějakou entitu či hodnotu, jednu anotaci či nabídku anotace, která koresponduje s nejpravděpodobnější alternativou. Uživatel pak může nabídku odmítnout (či smazat chybnou anotaci) a manuálně vytvořit správnou anotaci. Manuální vytváření anotace však může být velmi zdlouhavé – i při využití kontrolovaného slovníku je potřeba označit daný fragment, vyhledat příslušnou entitu ve slovníku apod., přičemž jsou všechny entity ze slovníku prezentovány stejně – tedy zcela nezávisle na anotovaném textu. Automatické anotační nástroje však často pracují ve dvou fázích, z nichž v první jsou vyhledány výskyty zmínek o entitách v textu a ve druhé je následně prováděno zjednoznačňování. Některé nástroje pak při zjednoznačňování určí míry důvěry ve správnost jednotlivých anotací a vybe-

rou tu nejlepší, která bude prezentována uživateli. Právě uvedené míry důvěry jsem využil pro sestavení seznamu alternativních nabídek anotací, jehož myšlenkou je usnadnit anotování v daných případech tím, že uživatel bude jedním kliknutím vybírat ze seznamu nejbližších alternativ a nikoliv z celého kontrolovaného slovníku.

Experimenty byly zaměřeny na to, zda navržený koncept prezentace alternativních nabídek anotací napomáhá při anotování, zejména při zjednodušování mnohoznačných jmen. Následně byl vyhodnocován vhodný způsob prezentace alternativ uživateli.

Experimenty prokázaly, že využití alternativ značně zvyšuje komfort uživatelů při anotování a zrychluje proces anotování. Výsledky testování různých způsobů prezentace alternativ uživateli sice obecně závisí na kvalitě procesu generování nabídek anotací v pozadí a na četnosti případů, kdy zmíněná entita není pokryta referenčními zdroji, ale okamžité zobrazení všech alternativ se ukazuje jako vhodnější, chceme-li v daném čase preferovat kvalitu vytvářených anotací před kvantitou. Naopak zobrazení alternativ až ve chvíli odmítnutí nejpravděpodobnější nabídky může mírně zrychlit proces anotování.

Ve druhé sadě experimentů pak bylo prokázáno, že když preferujeme kvalitu vytvářených anotací před kvantitou, může být výhodnější v alternativách prezentovat více informací. Nicméně jako celkově výhodnější se ukazuje zobrazení stručné zjednodušovací informace a možnost zobrazení detailů po kliknutí, neboť takové zobrazení snižuje náročnost zjednodušovací úlohy pro uživatele a zvyšuje rychlost procesu anotování, přičemž množství chyb narůstá pouze nepatrně. Pokud bychom chtěli dosáhnout znatelného zvýšení kvality, bylo by možné vynutit to, aby uživatel před konečným rozhodnutím prošel celý seznam alternativ, což se však ukázalo jako nepoužitelné řešení pro více než dvojnásobné zpomalení procesu anotování.

Rovněž se ukázalo, že uživatelé preferují možnost vracet se k nabídkám, které dříve vyloučili. Tedy i když mají možnost definitivního vyloučení dané nabídky, preferují pouze určité označení

vyloučených nabídek. Také se ukázalo, že různé výchozí zobrazení může vést na intuitivní využití jiného vzoru interakce uživatelem, neboť možnosti rozbalení detailních informací využilo menší procento uživatelů než možnosti jejich sbalení, přičemž sbalování intuitivně používali pro vylučování jednotlivých nabídek. Při rozbalování pak byl proces vylučování pomocí opětovného sbalení pozorovatelný výjimečně. Výchozí zobrazení a další možnosti interakce tedy mají zásadní vliv na činnost uživatele, což může vést na rozdílnou kvalitu anotací i na čas potřebný na splnění dané úlohy.

Nezávisle na způsobu zobrazení se však jasně prokázalo, že zobrazení alternativních nabídek anotací zvýší kvalitu vytvářených anotací i rychlost anotování, neboť sníží počet manuálně vyplňovaných anotací na nutné minimum a uživatelé primárně vybírají z menšího množství potenciálně korespondujících entit. Dotazníkové šetření rovněž prokázalo zvýšení komfortu pro uživatele.

7.6 Přínos sémantického filtrování

Nový koncept sémantického filtrování zahrnuje šablony pro jednotlivé sémantické typy anotací a asistenci uživateli při výběru hodnoty zvoleného atributu anotovaného konceptu. Šablony lze vytvořit importem ontologie, přičemž anotační systém před uživatelem skryje komplexitu ontologie a vizualizuje ji ve formě stromu anotací s atributy příslušných typů. Když uživatel např. vybere, že chce anotovat událost, jsou mu zobrazeny atributy vyplývající z popisu události v ontologii a jejich typy. Při výběru konkrétního atributu jsou pak v textu zvýrazněny entity či hodnoty korespondující s daným typem či jeho podtypy. Uživatel tak na první pohled vidí potenciální hodnoty daného atributu a může z nich vybrat jedním kliknutím. Není-li hodnota mezi rozpoznanými (zvýrazněnými), může ji doplnit manuálně.

Pro ověření přínosu sémantického filtrování byly provedeny dva experimenty. První se zaměřuje na přínos celkového přístupu k poloautomatickému anotování komplexních hierarchických anotací

využitého v nástroji 4A oproti manuálnímu přístupu. Zde je patrný značný pokles doby anotování (téměř na jednu třetinu) se současným zvýšením kvality vytvářených anotací (pokles chyb na méně než polovinu). Tím je prokázán jasný přínos využitého přístupu, ale není dostatečně určen příspěvek samotného sémantického filtrování.

Pouze na sémantické filtrování se proto zaměřil druhý provedený experiment, kde bylo srovnáno anotování událostí bez sémantického filtrování, kdy bylo jako hodnotu atributu možné vybrat jakoukoliv anotaci, přičemž žádná nebyla oproti ostatním zvýrazněna (byly pouze odlišeny barvami dle jednotlivých sémantických typů), s anotováním se zapnutým sémantickým filtrováním. Výsledky experimentu ukazují, že je proces anotování komplexních událostí pouze vlivem sémantického filtrování rychlejší o 15 %. Relativní snížení obou sledovaných typů chyb současně překročilo 25 %. Tím je jasně prokázáno, že sémantické filtrování zvýší rychlost a kvalitu vytvářených strukturovaných anotací.

7.7 Kolaborativní vytváření ontologie

Specializované nástroje na tvorbu ontologií jako je Protégé³ vyžadují obecnou znalost tvorby ontologií, zahrnující mimo jiné i porozumění základním konceptům jako jsou třída, individuál a literál, typy omezení apod. Znalostní inženýři, ovládající tyto oblasti, však obvykle nemají dostatečné znalosti z domény, kterou mají pomocí cílové ontologie popsat. Například u moderního průzkumu trhu pro správu reputace obchodních značek při tom specialista vytváří ontologie pro zákazníky z různých odvětví. V této situaci je nutná spolupráce se zákazníkem, který má vynikající znalosti z dané oblasti, ale obvykle nemá znalosti z oblasti tvorby ontologií ani zkušenosti s využitím příslušných nástrojů.

Spolupráce znalostního inženýra se zákazníkem může probíhat pouze nad samotnou vytvářenou ontologií vizualizovanou ve spe-

³<http://protege.stanford.edu/>

cializovaném nástroji. To však vyžaduje vzájemné vysvětlování, kdy znalostní inženýr zákazníkovi popisuje, jak chápe jednotlivé popisované koncepty, a zákazník jeho informace upřesňuje a doplňuje. Navržené řešení předpokládá spolupráci přímo nad texty, při jejichž analýze má být výsledná ontologie využita. Uživatel pak může navrhovanou ontologii přímo využít k anotování daného textu, čímž vzniknou případy použití umožňující lepší pochopení popisovaných konceptů. Mnou navržený anotační nástroj 4A současně poskytne odstínění uživatele od komplexity formálního popisu ontologie, a umožní tak jednoduchou spolupráci laika se znalostním inženýrem.

Provedený experiment prokázal, že pokud je ke spolupráci znalostního inženýra s informovaným klientem na tvorbě ontologie využit nástroj 4A, dojde k relativnímu zrychlení kolaborativního kroku tvorby ontologie o více než 20 %. Toto zrychlení je sice částečně kompenzováno mírným zpomalením individuální přípravy (zejména na straně znalostního inženýra), ale z vyhodnocení dotazníků po ukončení experimentu plyne, že konzultování ontologie nad reálnými anotovanými daty všichni zúčastnění vnímali jako zásadní vylepšení práce. Tím bylo prokázáno, že kolaborativní rozšiřování ontologie při anotování s využitím reálných anotovaných dat značně usnadňuje proces její tvorby a aktualizace.

7.8 Souhrnné výsledky

V rámci vývoje anotačního systému a provedených experimentů se podařilo splnit všechny stanovené hlavní cíle práce.

Vyvinutý systém splnil nejen požadavky stanovené v rámci této práce, ale i požadavky evropského projektu Decipher, kde usnadnil práci muzejních profesionálů při tvorbě výstav. Jeho kvality ukázalo i srovnání s existujícími nástroji, a to jak v teoretickém tabulovém přístupu, tak i při praktickém srovnávacím experimentu.

Při optimalizaci množství zobrazovaných informací se podařilo docílit zvýšení rychlosti anotování i kvality vytvářených anotací.

Navržené koncepty zobrazení alternativních nabídek anotací a sémantického filtrování pak vedly k dalšímu zlepšení.

Kolaborativní přístup k vytváření ontologie nad reálnými anotovanými daty pak nejen ocenili nejen účastníci experimentu, ale umožnil i dosažení efektivnější spolupráce, a tím i kratšího času v kolaborativním kroku experimentu.

Do budoucna se zaměřím na další vývoj anotačního systému a detailnější experimenty vycházející z vyhodnocení výsledků z jednotlivých kroků realizovaných v této práci. Inspirací pro budoucí vývoj anotačního systému bude např. editor Aloha⁴ umožňující editaci přímo v zobrazené webové stránce a v plánu je rovněž nový anotační doplněk do webového prohlížeče Google Chrome⁵. V praktických experimentech bude třeba blíže prozkoumat vliv zobrazované míry důvěry ve správnost jednotlivých alternativních nabídek anotací, detaily ve vzorech interakce uživatele při zjednotňování entit (možnost označování či klasifikace alternativních nabídek při procházení seznamu) a další detaily, které by mohly vést k dalším zlepšením procesu anotování a zjednodušení a zpříjemnění této úlohy pro uživatele. Chystám rovněž experimenty na paralelní kolaborativní práci v různých prostředích a vyhodnotím při ní příslušné vzory interakce na týmových úlohách.

⁴<http://www.alohaeditor.org/demo/aloha-ui/>

⁵<https://www.google.com/chrome/browser/desktop/index.html>

Kapitola 8

Závěr

V rámci této disertační práce jsem nejprve prostudoval teorii a existující řešení v oblasti anotování textu. Pod pojmem anotace jsem se rozhodl chápat doplňující informaci přidanou k textu bez omezení významu této informace či charakteru anotované části textu. S tímto rozhodnutím souvisí i následné zaměření práce na vytváření komplexních hierarchických anotací bez omezení úrovně zanoření a zkoumání vzorů interakce uživatelů s pokročilými anotačními nástroji při jejich vytváření.

Protože se již po prvotním prozkoumání dostupných informací o existujících anotačních nástrojích ukázalo, že žádný z nich neposkytuje funkcionalitu potřebnou pro uvedený výzkum, v rámci práce jsem se rozhodl vytvořit nový anotační systém. Správnost tohoto rozhodnutí potvrdilo mimo jiné i následné srovnání s existujícími řešeními.

Anotační systém 4A vytvořený v rámci této práce má architekturu klient – server, kde klientskou stranu reprezentují rozšíření pro WYSIWYG editory textu v jazyce JavaScript a doplňky či rozšířeními pro webové prohlížeče. Primární zaměření na anotování obsahu na webu vyplynulo z trendu vedoucího k online službám a potřeby metadat pro naplnění vize sémantického webu. Vytvořený systém zahrnuje nejen vlastnosti implementované v řadě sou-

časných anotačních systémů, které byly identifikovány jako důležité na základě dostupných studií a vlastního výzkumu, ale i řadu nových konceptů i různých vylepšení existujících metod. Hlavními přínosy jsou nový koncept nabízení anotací s alternativními nabídkami, sémantické filtrování usnadňující vytváření strukturovaných anotací a práce s odběry anotací umožňující pohledy na anotovaný text z různých perspektiv v několika panelech prohlížeče. Za zmínku stojí rovněž nová knihovna pro aktualizaci anotovaných fragmentů umožňující pokročilou aktualizaci anotací po offline editaci textu, pokročilá synchronizace dokumentu a anotací pro kolaborativní práci na paralelním vytváření a anotování textu a nový koncept klonování dokumentu s anotacemi.

V rámci vývoje systému byl navržen nový protokol pro přenos anotací a dalších informací potřebných při činnosti pokročilých anotačních nástrojů. Vzhledem k tomu, že dle dostupných informací bylo jediným existujícím řešením k tomuto účelu API od Lee Feigenbauma z firmy IBM z roku 2004, které není široce využíváno, mohl by navržený protokol či koncepty v něm využité umožnit budoucí standardizaci v této oblasti a interoperabilitu různých anotačních nástrojů. Současně byl v rámci této práce navržen i nový formát anotace, který poskytl alternativu k paralelně vznikajícímu novému formátu Web Annotation (dříve Open Annotation) a umožnil vyhodnocení jeho výhod a nevýhod. Některé části specifikace formátu 4A by v budoucnu mohly být základem pro rozšíření formátu Web Annotation o šablony atributů pro jednotlivé sémantické typy anotací, které umožňují využití pokročilých technik pro vytváření hierarchických anotací.

Již prvotní testy systému 4A ukázaly, že podporuje objevování znalostí a pomáhá tak porozumět vědeckým tématům společnou prací na vytváření znalostních struktur.

Systém 4A byl úspěšně nasazen v evropském projektu Decipher, kde byl využit jeho potenciál pro zjednodušení tvorby hierarchických anotací. Muzejní profesionálové ocenili možnost vyhledávání komplexních událostí přímo v textu a jejich okamžitý

export do systému, ve kterém vytvářejí popisy výstav na základě historických příběhů a vyprávění. Pokročilé anotování usnadnilo i řadu dalších činností, které uživatelé v muzeích při přípravě výstav běžně provádějí.

Po teoreticky zaměřeném tabulkovém srovnání nového systému s existujícími nástroji byly provedeny praktické srovnávací experimenty s vybranými nástroji z těch, které ze srovnání vyšly jako nejpokročilejší. Ukázalo se, že 4A díky vhodné kombinaci známých i nových funkcí umožňuje dosáhnout lepších výsledků anotování v kratším čase. Rovněž bylo prokázáno, že je přínosné vizuální odlišení nabídek anotací generovaných systémem od anotací zkontrolovaných či vytvořených uživatelem, a projevil se i přínos nového konceptu sémantického filtrování v nástroji 4A. Výsledky nepochybně prokazují, že uživatelské rozhraní zásadně ovlivňuje nejen rychlost anotování, ale i kvalitu vytvářených anotací, čímž byl splněn jeden z cílů méj práce.

Nástroj 4A byl navržen nejen jako lepší řešení pro anotování textu, ale také jako platforma pro testování nových přístupů v oblasti anotování. Pro ověření řady konceptů totiž historicky vznikaly samostatné nástroje, jejichž primární zaměření na ověření daného konceptu často vedlo k omezením zabraňujícím širšímu využití a v řadě případů i k následnému ukončení vývoje i podpory daného nástroje. Jedním z mých cílů proto bylo vytvořit nástroj, který díky své modulární architektuře a širokým možnostem konfigurace umožní detailní ověřování různých konceptů při vynaložení minimálního úsilí na vývoj potřebného programového vybavení. Následné experimenty ukázaly, že se tento cíl podařilo splnit.

V následujících experimentech jsem se zaměřil na zkoumání vzorů interakce uživatelů s anotačním nástrojem a možností zrychlení procesu anotování a zlepšení kvality vytvářených anotací.

První pokročilé experimenty byly zaměřeny na optimalizaci množství zobrazovaných informací pro zrychlení procesu zjednotňování víceznačných jmen entit a zvýšení kvality vytvářených anotací. Výsledky ukazují, že často využívaný přístup, při kterém

je zobrazen pouze typ entity a URL na stránku popisující danou entitu v referenčním zdroji, vede na největší chybovost i nejdelší čas anotování. Zobrazení detailních informací přímo u nabídky anotace však rovněž nevede k nejlepším výsledkům. Ve srovnání je zde sice relativní pokles množství chyb o 78 %, stále se však jedná o nezanedbatelné množství. Chyby jsou často způsobeny tím, že velké množství zobrazených informací vede k rozptýlení uživatele a rychlejší únava ze čtení většího množství textu snižuje jeho pozornost. Jako nejvhodnější se ukázalo zobrazení stručného zjednotňovacího atributu a možnost následného zobrazení detailních informací dle potřeby. Oproti přímému zobrazení detailních informací je zde relativní pokles množství chyb o více než 66 %, oproti nejčastějšímu zobrazení typu a URL o více než 92 %. Současně se oproti zobrazení typu a URL snížil průměrný čas na anotaci o 11,5 % a četnost potřeby návštěvy referenčního zdroje o 96 %. Experiment prokázal, že zvolené optimální množství zobrazovaných informací tedy zvýšilo rychlost anotování i kvalitu vytvářených anotací, čímž byl splněn další z cílů mé práce. Jednoduchý algoritmus pro sestavení obsahu zjednotňovacího atributu by do budoucna bylo možné dále vylepšit o kontextově závislé informace, a potenciálně tak dosáhnout dalšího zlepšení. Toto bude jedním z témat mého budoucího výzkumu v oblasti zjednotňování entit.

Druhá sada experimentů byla zaměřena na nový koncept zobrazování alternativních nabídek anotací a vhodný způsob jejich prezentace uživateli. Byl srovnán přístup, kdy jsou alternativy zobrazeny po odmítnutí nejpravděpodobnější nabídky či na vyžádání, s přístupem, kdy jsou ihned zobrazeny všechny alternativy. Zobrazení všech alternativ se ukázalo jako vhodnější, chceme-li v daném čase preferovat kvalitu vytvářených anotací před kvantitou. Nicméně současný nárůst průměrného času na anotaci vedl k nepřesvědčivým výsledkům z pohledu případného celkového zlepšení. Zcela jistě však mají oba způsoby prezentace znatelný vliv na způ-

sob interakce uživatele a vedou k jiné intuitivní realizaci kompromisu mezi rychlostí a kvalitou vytvářených anotací.

Vzhledem k tomu, že bez zobrazení alternativ by nutně muselo dojít k dalšímu zpomalení, neboť manuální tvorba anotace se už v prvotních experimentech ukázala jako velmi náročná a byla by potřebná vždy, když nejpravděpodobnější nabídka není správná, což bylo v provedených experimentech ve 40 % (dle empirického měření přesnosti nástroje pro automatické generování nabídek anotací), je současně prokázáno, že zobrazení alternativních nabídek anotací zvýší rychlost anotování i kvalitu vytvářených strukturovaných anotací. Způsob prezentace pak pouze ovlivní úroveň pro kompromis mezi kvalitou a kvantitou.

Třetí sada experimentů ověřovala přínos nabízení anotací se sémantickým filtrováním při anotování komplexních událostí. Relativní pokles počtu nesprávných hodnot o 61,5 %, chybějících hodnot o 48,5 % a času na anotování události o 64 % jasně ukazují výrazný přínos konceptu nabízení anotací využitého v systému 4A. Nabízení anotací a sémantické filtrování tedy prokazatelně zvyšují rychlost a kvalitu vytvářených strukturovaných anotací.

Čtvrtá sada experimentů pak zjišťovala přínos samotného konceptu sémantického filtrování. Relativní snížení počtu chyb o 26 % a času na anotování vztahu o 15 % prokazují, že přínos sémantického filtrování je nezanedbatelný.

Pátá sada experimentů byla zaměřena na rozdíl mezi okamžitou prezentací detailních informací o entitě, včetně stručné zjednotčovací informace, a stručnou prezentací umožňující zobrazení detailů. Tímto byl rozšířen předchozí experiment na optimalizaci zobrazovaných informací a bylo ověřeno, že rozdíly nejsou způsobeny pouze faktem, že je zjednotčovací atribut uveden. Při okamžitém zobrazení všech informací dojde oproti stručnému zobrazení k relativnímu snížení počtu chyb o 6 %, ale také k nárůstu času anotování o 35 % a ke značnému snížení komfortu uživatelů, pro které se tak úloha stane výrazně náročnější. Stručné zobrazení s možností rozbalení detailních informací se tedy ukazuje ekono-

micky výhodnější alternativou, zatímco detailní zobrazení umožňuje zvýšení kvality anotací. Dalšího zvýšení kvality by pak bylo možné dosáhnout vynucením procházení všech alternativ, které by vedlo na více než dvojnásobný nárůst času, což se zdá ekonomicky nepřijatelné.

Poslední prováděný experiment prokázal, že kolaborativní rozšiřování ontologie nad reálnými daty anotovanými právě vytvářenou ontologií vede k usnadnění a zrychlení tohoto procesu. Tím bylo současně dokázáno, že strukturované anotace umožňují lepší popis sémantiky, přičemž pokud tyto anotace využijeme při kolaborativním anotování v reálném čase, usnadní se tím spolupráce při tvorbě ontologií.

V budoucím výzkumu se zaměřím na zkoumání vlivu zobrazené míry důvěry ve správnost nabídky anotace na rychlost a správnost rozhodování uživatelů, optimalizaci počtu zobrazených alternativních nabídek anotací (tedy práh míry důvěry pro zobrazení alternativy), možnosti vylepšení přístupu ke kontrolovanému slovníku, kolaborativní paralelní anotování větším množstvím uživatelů a na další pokročilé přístupy k anotování textu.

Literatura

- [1] Agosti, M.; Albrechtsen, H.; Ferro, N.; aj.: DiLAS: a Digital Library Annotation Service. In *Proceedings of Annotation for Collaboration – A Workshop on Annotation Models, Tools and Practices*, 2005, [Online; navštíveno 19. 1. 2011]. URL http://www.is.informatik.uni-duisburg.de/bib/pdf/ir/Agosti_etal_05.pdf
- [2] Agosti, M.; Bonfiglio-Dosio, G.; Ferro, N.: A historical and contemporary study on annotations to derive key features for systems design. *International Journal on Digital Libraries*, ročník 8, 2007: s. 1–19, ISSN 1432-5012, [Online; navštíveno 19. 1. 2011]. URL <http://www.springerlink.com/content/93j367635270641h/fulltext.pdf>
- [3] Agosti, M.; Ferro, N.: A System Architecture as a Support to a Flexible Annotation Service. In *Peer-to-Peer, Grid, and Service-Oriented in Digital Library Architectures, Lecture Notes in Computer Science*, ročník 3664, editace C. Türker; M. Agosti; H.-J. Schek, Springer Berlin / Heidelberg, 2005, s. 147–166, [Online; navštíveno 19. 1. 2011]. URL <http://www.springerlink.com/content/90xvqnnwajaktd7l/fulltext.pdf>
- [4] Bottoni, P.; Levialdi, S.; Rizzo, P.: An Analysis and Case Study of Digital Annotation. In *Databases in Networked*

Information Systems, Lecture Notes in Computer Science, ročník 2822, editace N. Bianchi-Berthouze, Springer Berlin / Heidelberg, 2003, s. 216–231, [Online; navštíveno 19. 1. 2011].

URL <http://www.springerlink.com/content/u1cu2mx7cexjf01h/fulltext.pdf>

- [5] Ciccarese, P.; Ocana, M.; Clark, T.: Open semantic annotation of scientific publications using DOME0. *Journal of Biomedical Semantics*, ročník 3, č. Suppl 1, January 2012, [Online; navštíveno 2. 8. 2016].
URL <http://www.jbiomedsem.com/content/3/S1/S1>
- [6] Cunningham, H.; Maynard, D.; Bontcheva, K.; aj.: *Text Processing with GATE (Version 6)*. GATE, 2011, ISBN 978-0956599315, [Online; navštíveno 2. 8. 2016].
URL <http://tinyurl.com/gatebook>
- [7] Dytrych, J.; Smrz, P.: Interaction Patterns in Computer-Assisted Semantic Annotation of Text – An Empirical Evaluation. In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence*, Rome, Italy: SCITEPRESS, 2016, ISBN 978-989-758-172-4.
URL <http://www.scitepress.org/DigitalLibrary/PublicationsDetail.aspx?ID=5F3T/EA+bgg&t=1>
- [8] Dytrych, J.; Smrz, P.: *Springer Book of ICAART 2016*, kapitola Advanced User Interfaces for Semantic Annotation of Complex Relations in Text. *Lecture Notes in Computer Science*, Springer International Publishing, 2016.
- [9] Feigenbaum, L.: Standardize annotations with Web services. IBM, Duben 2004, [Online; navštíveno 17. 12. 2010].
URL <http://www.ibm.com/developerworks/webservices/library/ws-annotation.html>

- [10] GATE Developer. The University of Sheffield, 2015, [Online; navštíveno 2. 12. 2015].
URL <https://gate.ac.uk/family/developer.html>
- [11] Hinze, A.; Heese, R.; Luczak-Rösch, M.; aj.: Semantic enrichment by non-experts: usability of manual annotation tools. In *Proceedings of the 11th international conference on The Semantic Web*, ročník Part I, Springer-Verlag Berlin Heidelberg, 2012, s. 165–181, [Online; navštíveno 2. 8. 2016].
URL <http://iswc2012.semanticweb.org/sites/default/files/76490161.pdf>
- [12] Khalili, A.: Annozilla (Annotea on Mozilla). 2015, [Online; navštíveno 29. 12. 2015].
URL <http://rdface.aksw.org/>
- [13] Khalili, A.; Auer, S.; Hladky, D.: The RDFa Content Editor – From WYSIWYG to WYSIWYM. In *Proceedings of COMPSAC 2012 – Trustworthy Software Systems for the Digital Society*, 2012, [Online; navštíveno 2. 8. 2016].
URL http://svn.aksw.org/papers/2012/COMPSAC2012_RDFaCE/public.pdf
- [14] LaLiberte, D.; Braverman, A.: A protocol for scalable group and public annotations. *Comput. Netw. ISDN Syst.*, ročník 27, Duben 1995: s. 911–918, ISSN 0169-7552, [Online; navštíveno 19. 1. 2011].
URL http://ac.els-cdn.com/0169755295000172/1-s2.0-0169755295000172-main.pdf?_tid=6c512aae-526b-11e6-b535-00000aacb35f&acdnat=1469453381_8adf36c91b925390962ad4a8560a43e0
- [15] Maynard, D.: Benchmarking Textual Annotation Tools for the Semantic Web. In *6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco: European Language Resources Association (ELRA), 2008,

[Online; navštíveno 2. 8. 2016].

URL

<https://gate.ac.uk/sale/lrec2008/benchmarking.pdf>

- [16] Maynard, D.; Dasiopoulou, S.; Costache, S.; aj.: Knowledge Web Project: Deliverable D1.2.2.1.3 – Benchmarking of annotation tools. 2007, [Online; navštíveno 2. 8. 2016].

URL <http://knowledgeweb.semanticweb.org/semanticportal/deliverables/D1.2.2.1.3.pdf>

- [17] Moro, A.; Navigli, R.: SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, Denver, Colorado, June 2015, s. 288–297, [Online; navštíveno 2. 8. 2016].

URL <http://www.aclweb.org/anthology/S15-2049>

- [18] TinyMCE. Moxiecode Systems AB, 2011, [Online; navštíveno 2. 1. 2011].

URL <http://tinymce.moxiecode.com/>

- [19] Niranatlamphong, W.; Choochaiwattana, W.; Spring, M.: A conceptual framework for digital annotation system on WWW. In *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*, Srpen 2009, s. 27–31, [Online; navštíveno 17. 1. 2011].

URL <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5234462>

- [20] Phelps, T.; Wilensky, R.: Multivalent Annotations. In *Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science*, ročník 1324, editace C. Peters; C. Thanos, Springer Berlin / Heidelberg, 1997, s. 287–303, [Online; navštíveno 19. 1. 2011].

URL <http://www.springerlink.com/content/d87v173g7671j20p/fulltext.pdf>

- [21] da Rocha, T.; Willrich, R.; Fileto, R.; aj.: Supporting Collaborative Learning Activities with a Digital Library and Annotations. In *Education and Technology for a Better World, IFIP Advances in Information and Communication Technology*, ročník 302, editace A. Tatnall; A. Jones, Springer Boston, 2009, s. 349–358, [Online; navštíveno 19. 1. 2011].
URL <http://www.springerlink.com/content/gug76w637102x821/fulltext.pdf>
- [22] Sanderson, R.; Ciccarese, P.; de Sompel, H. V.; aj.: Open Annotation Data Model. Open Annotation Community Group, 2013, [Online; navštíveno 5. 1. 2016].
URL <http://www.openannotation.org/spec/core/>
- [23] Sanderson, R.; Ciccarese, P.; Young, B.; aj.: Web Annotation Data Model. W3C Candidate Recommendation, Červenec 2016, [Online; navštíveno 1. 8. 2016].
URL <https://www.w3.org/TR/annotation-model/>
- [24] Schickler, M. A.; Mazer, M. S.; Brooks, C.: Pan-browser support for annotations and other meta-information on the World Wide Web. In *Proceedings of the fifth international World Wide Web conference on Computer networks and ISDN systems*, Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 1996, s. 1063–1074, [Online; navštíveno 2. 8. 2016].
URL <http://www.sciencedirect.com/science/article/pii/0169755296000608>
- [25] SharedCopy. SharedCopy, 2010, [Online; navštíveno 29. 12. 2010].
URL <http://sharedcopy.com/>
- [26] Smrz, P.; Dytrych, J.: Towards New Scholarly Communication: A Case Study of the 4A Framework. In

SePublica, CEUR Workshop Proceedings, ročník 721, Ruzica Piskac, Redaktion Sun SITE, Informatik V, RWTH Aachen, Ahornstr. 55, 52056 Aachen, Germany, 2011, ISSN 1613-0073, [Online; navštíveno 2. 8. 2016].
URL <http://ceur-ws.org/Vol-721/paper-07.pdf>

- [27] Smrz, P.; Dytrych, J.: Advanced Features of Collaborative Semantic Annotators – the 4A System. In *Proceedings of the 28th International FLAIRS Conference*, Hollywood, Florida, USA: AAAI Press, 2015, [Online; navštíveno 2. 8. 2016].
URL <http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS15/paper/view/10356>
- [28] Stenetorp, P.; Pyysalo, S.; Topić, G.; aj.: BRAT: A Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, s. 102–107, [Online; navštíveno 19. 7. 2016].
URL <http://dl.acm.org/citation.cfm?id=2380921.2380942>
- [29] Surdeanu, M.; Heng, J.: Overview of the English Slot Filling Track at the TAC2014 Knowledge Base Population Evaluation. In *Proceedings of the TAC-KBP 2014 Workshop*, 2013=4, [Online; navštíveno 2. 8. 2016].
URL <http://nlp.cs.rpi.edu/paper/sf2014overview.pdf>
- [30] Völkel, M.; Haller, H.: Conceptual Data Structures (CDS) – Towards an Ontology for Semi-Formal Articulation of Personal Knowledge. In *14th International Conference on Conceptual Structures*, Aalborg University, Denmark, 16.07.06, 2006, [Online; navštíveno 17. 12. 2010].
URL <http://www.xam.de/2006/04-cds.pdf>

- [31] Annotea Project. W3C, 2010, [Online; navštíveno 29. 12. 2010].
URL <http://www.w3.org/2001/Annotea/>
- [32] Weng, C.; Gennari, J. H.: Asynchronous collaborative writing through annotations. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work, CSCW '04*, New York, NY, USA: ACM, 2004, ISBN 1-58113-810-5, s. 578–581, [Online; navštíveno 19. 1. 2011].
URL http://portal.acm.org/ft_gateway.cfm?id=1031705&type=pdf&CFID=6206112&CFTOKEN=33710793
- [33] Yee, K. P.: CritLink: Advanced Hyperlinks Enable Public Annotation on the Web. 2002,
<http://zesty.ca/pubs/cscw-2002-crit.pdf>.
URL <http://zesty.ca/pubs/cscw-2002-crit.pdf>
- [34] Zheng, Q.; Booth, K.; McGrenere, J.: Co-Authoring with Structured Annotations. Department of Computer Science, University of British Columbia, 2006, [Online; navštíveno 17. 12. 2010].
URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.2558&rep=rep1&type=pdf>

Přílohy

Seznam příloh

A	Strukturovaný životopis	69
B	Seznam publikací	73
C	Seznam inženýrských děl	75

Příloha A

Strukturovaný životopis

Osobní data

Jméno a příjmení: Ing. Jaroslav Dytrych
Datum narození: 31. 12. 1984
Místo narození: Třebíč
Státní příslušnost: ČR
Trvalé bydliště: Tetčická 663, 665 01 Rosice

Vzdělání

- 2000 – 2004** Střední průmyslová škola elektrotechnická, Kounicova 16, Brno
- v červnu 2004 maturita v oboru Slaboproudá elektrotechnika – Elektronické počítačové systémy
- 2004 – 2007** bakalářský studijní program IT-BC-3 na Fakultě informačních technologií VUT v Brně
- v červnu 2007 státní závěrečná zkouška v oboru Informační technologie

2007 – 2009 magisterský studijní program IT-MGR-2 na Fakultě informačních technologií VUT v Brně

– v červnu 2009 státní závěrečná zkouška v oboru
Informační systémy

od 2009 doktorský studijní program VTI-DR-4 na Fakultě informačních technologií VUT v Brně

Pracovní zkušenosti a odborná praxe

- červen 2002 – 2 týdny odborné praxe ve firmě Aicom Brno
- červen 2003 – 2 týdny odborné praxe ve firmě Aicom Brno
- od roku 2004 správce webu firmy Zahradnictví Libor Gotwald (www.upravy-zahrad.com)
- od roku 2004 správce webu firmy Dagmar Halasová – Sběratelské předměty (www.sberatelske-predmety.com)
- od roku 2006 správcem počítačové sítě Správy zámku Rosice
- od roku 2008 správce webu Správy zámku Rosice (www.zamek.rosice.cz)
- od roku 2009 správce informačního systému výzkumné skupiny NLP@FIT (nyní KNOT@FIT) na VUT v Brně
- od roku 2010 výuka v předmětu Praktické aspekty vývoje software (IVS) – organizace projektů a přednášky o týmové spolupráci a sdílení dat v týmu
- v letech 2011 – 2014 účast na evropském projektu Decipher zkoumajícím kulturní dědictví prostřednictvím příběhů a vyprávění – <http://decipher-research.eu/> (od r. 2012 zodpovědný za vývoj v projektovém týmu z VUT)

- od roku 2012 správce serverů výzkumné skupiny KNOT (v roce 2016 již >50 serverů)
- v roce 2012 měsíční stáž na OpenUniversity v Anglii v rámci projektu Decipher
- od roku 2014 zástupce studijního poradce FIT VUT v Brně
- v letech 2015 a 2016 jsem poskytoval technickou podporu kolegům v rámci účasti na evropském projektu JUNIPER (<http://www.juniper-project.org/>)
- v letech 2015 a 2016 účast na evropském projektu MixedEmotions (<http://mixedemotions-project.eu/>)
- od roku 2016 výuka (přednášky) v předmětu Grafická uživatelská rozhraní v Javě (GJA)
- od roku 2016 poradce pro studijní agendy v informačním systému FIT VUT v Brně
- od roku 2016 účast na projektu NAKI II CPK – Využití sémantických technologií pro zpřístupnění kulturního dědictví prostřednictvím centrálního portálu knihoven

Odborná orientace

- Pokročilé webové technologie
- Informační systémy
- Počítačové sítě

Další znalosti a dovednosti

- Angličtina: středně pokročilý
- Řidičský průkaz: B

Ostatní zájmy

- elektronika
- tanec
- plavání

Příloha B

Seznam publikací

- Dytrych, J.; Smrz, P.: Interaction Patterns in Computer-Assisted Semantic Annotation of Text – An Empirical Evaluation. In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence*, Rome, Italy: SCITEPRESS, 2016, ISBN 978-989-758-172-4. URL <http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0005695900740084>
- Smrz, P.; Dytrych, J.: Towards New Scholarly Communication: A Case Study of the 4A Framework. In *SePublica, CEUR Workshop Proceedings*, ročník 721, Ruzica Piskac, Redaktion Sun SITE, Informatik V, RWTH Aachen, Ahornstr. 55, 52056 Aachen, Germany, 2011, ISSN 1613-0073. URL <http://ceur-ws.org/Vol-721/paper-07.pdf>
- Smrz, P.; Dytrych, J.: Advanced Features of Collaborative Semantic Annotators – the 4A System. In *Proceedings of the 28th International FLAIRS Conference*, Hollywood, Florida, USA: AAAI Press, 2015. URL <http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS15/paper/view/10356>

Přijato k publikování

- Dytrych, J.; Smrz, P.: *Springer Book of ICAART 2016*, kapitola Advanced User Interfaces for Semantic Annotation of Complex Relations in Text. Lecture Notes in Computer Science, Springer International Publishing, 2016.

Spoluautorství na technických zprávách

- Smrz, P.; Otrusina, L.; Dytrych, J.; aj.: DECIPHER Deliverable D4.2.1: Relationship Mining Component. Technická zpráva, Brno University of Technology, 2012.
- Smrz, P.; Otrusina, L.; Dytrych, J.; aj.: DECIPHER Deliverable D4.3.1: DECIPHER Semantic Annotator. Technická zpráva, Brno University of Technology, 2013.
URL <http://cordis.europa.eu/docs/projects/cnect/1/270001/080/deliverables/001-DecipherD431SemanticAnnotatorv01.pdf>

Příloha C

Seznam inženýrských děl

- 2010** Server pro sdílení anotací, software, 2010. Autoři: Dytrych Jaroslav, Smrž Pavel URL <http://www.fit.vutbr.cz/research/prod/index.php?id=176>
- 2012** Editor anotací, software, 2012. Autoři: Cudrák Miloš, Dytrych Jaroslav, Kleban Martin, Loukota Petr, Pěnkava Pavel, Smrž Pavel URL <http://www.fit.vutbr.cz/research/prod/index.php?id=272>
- 2013** Anotační systém pro Drupal, software, 2013. Autoři: Cudrák Miloš, Dytrych Jaroslav, Kleban Martin, Kopecký Marek, Kratochvílová Marie, Kubík Lukáš, Loskot Radim, Loukota Petr, Macháček Jakub, Pěnkava Pavel, Petr Martin, Smrž Pavel, Strecha Juraj, Žurek Aleš URL <http://www.fit.vutbr.cz/research/prod/index.php?id=328>
- 2015** Programy pro zpracování korpusů, software, 2015. Autoři: Doležal Jan, Dytrych Jaroslav, Karásek Miroslav, Kouřil Jan, Otrusina Lubomír, Smrž Pavel URL <http://www.fit.vutbr.cz/research/prod/index.php?id=482>

2015 Komponent pro sémantické obohacování textu, software,
2015 Autoři: Doležal Jan, Dytrych Jaroslav, Kouřil Jan,
Otrusina Lubomír, Smrž Pavel URL <http://www.fit.vutbr.cz/research/prod/index.php?id=479>

