Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Ján Faryad
Název práce	Rozpoznávání koreference pro Universal Dependencies
Rok odevzdání	2018
Studijní program	Informatika
Studijní obor	Obecná informatika
Autor posudku	Rudolf Rosa Oponent
Pracoviště	Ústav formální a aplikované lingvistiky

K celé práci

	lepší	OK	horší	nevyhovuje
Obtížnost zadání	X			
Splnění zadání		X		
Rozsah práce *... textová i implementační část, zohlednění náročnosti*	X			

Zadání práce je poměrně rozsáhlé – zahrnuje 4 body, z nichž každý vyžaduje nejprve seznámení s daným tématem nad rámec (či zcela mimo rámec) předmětů vyučovaných v bakalářském a částečně i magisterském studiu na MFF UK, a následně netriviální množství vlastní práce jak teoretické (návrh anotačního schématu, návrh systému strojového učení pro určování koreference, návrh evaluační míry), tak praktické (implementace konverze anotovaných dat do navrženého formátu, implementace a evaluace strojového učení).

Při plnění jednotlivých bodů se student často dopustil různých zjednodušení či ne zcela vhodných kroků, ale v rámci takto rozsáhlé bakalářské práce to považuji za pochopitelné a přijatelné. Dá se tedy konstatovat, že student uspokojivě splnil celé zadání, což u takto náročného zadání hodnotím jako práci nadprůměrně rozsáhlou. Dokonce poněkud nad rámec zadání student vyzkoušel několik různých algoritmů strojového učení a provedl stručnou diskuzi nad jejich výsledky, čímž motivoval finální volbu metody (bohužel je tato část práce poměrně stručná a nezahrnuje dílčí výsledky, student uvádí pouze evaluaci metody zvolené ve finální podobě práce).

Textová část práce

	lepší	OK	horší	nevyhovuje
Formální úprava *... jazyková úroveň, typografická úroveň, citace*	X			
Struktura textu *... kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu*			X	
Analýza			X	
Vývojová dokumentace		X		
Uživatelská dokumentace	X			

Práce je psána čtivě, dobrou češtinou, s naprostým minimem chyb a překlepů, a s vhodně užitou terminologií. Výslovně oceňuji obrázek 4, kde je velmi přehledně vizualizován formát CoNLL-U; ačkoliv s tímto formátem již několik let pracuji, takto přehledně jsem jej ještě zobrazen neviděl! Za jediné závažnější chybné místo, co se týče překlepů, považuji stranu 31, kde si autor v tabulce a následně i v ukázkovém výpočtu opakovaně plete precision a recall, takže čísla v tabulce i ve výpočtech jsou chybná. Vzhledem k tomu, že výpočet je uvedený pro ilustraci, tak zejména ilustruje buď to, že se student v problematice dobře neorientuje, anebo že práce byla psána ve spěchu; na druhou stranu jde pouze o ilustrační příklad, který tak nemá vliv na skutečnou evaluaci uvedenou v práci (ve zdrojových kódech se evaluace zdá být implementována korektně). Přitom metodu evaluace jako takovou považuji za dobře navrženou.

Při plnění prvního bodu (návrh anotace koreference v UD) práce zcela opomíjí skutečnost, že tento problém již do značné míry řeší návrh Enhanced UD (Schuster a Manning, 2016, LREC), který přímo zavádí anotaci vybraných podtypů koreference, a dle kterého se v současnosti již v některých jazycích tyto anotace provádějí. Toto opomenutí je o to překvapivější, že vedoucí práce se osobně účastnil konference, na které byl tento návrh představen. Student naproti tomu tvrdí, že v UD chybí možnost vyznačení koreference, a to přesto, že mu byla bakalářská práce zadána rok po publikování výše uvedeného článku, a dokončena o další rok později, kdy již jsou v rámci UD přímo k dispozici data s touto anotací. Přitom způsob anotace, který student navrhuje, má oproti oficiální anotaci v UD určité výhody, které by student mohl v práci diskutovat a svůj návrh tím vhodně motivovat (zejména jde o koreferenci napříč větami, ale i o koreferenci vypuštěných podmětů). Student možnosti Enhanced UD dokonce zmiňuje (v kontextu anotace přidaných umělých uzlů), takže o jejich existenci jako takové zjevně ví.

Za podstatný problém navržené metody anotace považuji, že předpokládá jen vypouštění zájmen v postavení podmětu, přestože student správně uvádí, že existují i jazyky, kde mohou být vypuštěna i jiná zájmena (kromě zmiňované japonštiny je to například také čínština, svahilština a mnohé původní americké jazyky); přitom v zadání práce je explicitně uvedeno, že navržené řešení má klást důraz na jazykovou univerzálnost. Logické by proto bylo zvolit druhou ze zvažovaných možností anotace, tj. pomocí doplnění umělých uzlů, kde by tento problém nebyl. Jak práce uvádí, v UD již tento koncept existuje pro nevyjádřené přísudky, rozšíření na nevyjádřená zájmena by tedy bylo logické, přímočaré, a poměrně čisté. Z textu práce to takto působí, že autor zvážil obě možnosti, a pak si z nich překvapivě zvolil tu zjevně méně vhodnou. Je také možné, že autor zvolil způsob anotace bez dostatečného zvážení jeho nevýhod, a ve chvíli, kdy zjistil, že by bývalo bylo lepší zvolit jiný způsob, tak již bylo pozdě – pak by to ale takto mělo být v práci uvedeno: nejprve analýza, která předcházela volbě způsobu anotace, pak volba anotace, a následně diskuze výhod a nevýhod zvolené anotace, třeba včetně poučení do budoucna v podobě upraveného návrhu anotace. To, že se provede analýza, na jejím základě se zvolí způsob anotace, a až během či po dokončení anotace nad reálnými daty se ukáže, že způsob anotace není zcela vhodný, je zcela běžné a je to v pořádku (je to spíše pravidlem než výjimkou). To, co není v pořádku, je provést správnou analýzu, ale pak si v rozporu s ní zvolit nevhodné řešení.

(pokračování na další straně)

Metodologicky považuji za překvapivé, že při převodu anotace koreference z PDT do UD autor nevyužívá již existující převod morfologické a syntaktické anotace PDT do UD, ale místo toho morfologickou a syntaktickou anotaci získává automatickou analýzou textu pomocí nástroje UDPipe. Obecně je takový postup samozřejmě v pořádku, nelze obecně předpokládat, že datové zdroje s anotací koreference jsou již převedené do UD, a autor správně uvádí, že kompletní převod do UD je velmi náročný. V případě PDT se ale využití existující konverze přímo nabízí. Autor uvádí, že využití automatické analýzy je vhodnější pro následné trénování nástroje pro detekci koreference, neboť je nutné předpokládat, že tento nástroj bude aplikován na automaticky analyzovaná data. Tento argument je sice dobrý a pravdivý, ale přesto závěr o nevhodnosti využití zlaté anotace nelze takto jednoduše učinit bez toho, aby byl podložen experimenty (v praxi se ukazuje, že někdy je vhodnější využít automatickou anotaci, a jindy zase zlatou anotaci, a není vždy zřejmé, proč tomu tak v konkrétních případech je).

Ne zcela šťastně jsou řešeny situace, kdy syntaktická anotace PDT značí koreferenci na uzlu, který je kořenem struktury v PDT, ale v UD kořenem této struktury není (autor výslovně zmiňuje koordinace, velice pravděpodobně existují i další takové případy). Jelikož i v autorem navrhované anotaci se koreference značí pouze u kořene struktury, naivní převod těchto struktur dopadá špatně, což autor pozoruje, ale neřeší. Přitom se nabízí anotaci koreference v PDT nejprve v podstatě převést do stylu OntoNotes, tj. vyznačit ji na všech uzlech v podstromu označeného kořene, přenést do UD anotace, a následně použít obdobný postup jako u OntoNotes, tj. nalézt (již v UD anotaci) kořen koreferující skupiny, a koreferenci vyznačit na něm. Samozřejmě zde pak může nastat problém, který nastává u OntoNotes, kdy v analýze z UDPipe koreferenční celek někdy netvoří souvislý podstrom. Tyto případy autor řeší tak, že anotaci koreference zahodí; lépe by bylo použít nějakou heuristiku, například nalézt největší podstrom patřící do koreferenčního celku, a koreferenci vyznačit na jeho kořeni.

Na straně 42 autor nepravdivě tvrdí, že data z korpusů nejsou volně k dispozici; ve skutečnosti toto platí pouze pro OntoNotes, zatímco PDT je volně ke stažení pod licencí CC-BY-NC-CA.

Není mi zcela jasná definice příznaků užitých ve strojovém učení, popis zde není dostatečně podrobný. Nejsem si také jistý jazykovou univerzálností specifického přístupu k různým typům zájmen; přitom většina přístupu je pro všechna zájmena stejná, není tedy zřejmé, zda je potřeba skutečně různá zájmena řešit takto odlišně. Ostatně, pokud je typ zájmene součástí použité sady příznaků, pak by případné rozhodnutí o užitečnosti či neužitečnosti konkrétních příznaků pro konkrétní typy zájmen měl být algoritmus strojového učení schopen provést sám od sebe.

Obecně je textová část práce místy příliš stručná. Asi nejvýraznější je tento nedostatek v případě kapitoly 5, popisující experimenty s automatickým určováním koreference, která má pouze necelé dvě strany a detailněji popisuje pouze jeden experiment, byť z textu vyplývá, že jich student provedl větší množství. Dílčí experimenty a jejich výsledky jsou jen velice vágně naznačeny několika slovy, a i hlavní experiment je vyhodnocen jen poměrně stručně, bez detailnější chybové analýzy. Je to škoda, z vědeckého pohledu jde přitom o potenciálně nejzajímavější část práce. Stejně tak chybí jakékolik porovnání s úspěšností jiných autorů řešících stejný problém.

Teoretické části práce jsou také spíše stručné, nerozebírají problém dostatečně do hloubky a ukazují jen základní orientaci studenta v dané problematice. To je ostatně patrné i z toho, že student cituje pouze 9 prací, z nichž je navíc zhruba polovina spíše technického rázu. Předloženou práci je tedy nutno nazírat zejména jako technickou a implementační, nikoli jako výzkumnou, byť tento potenciál má. Jelikož se ale jedná o práci bakalářskou, lze ji, s ohledem na formulaci jejího zadání, i přes tento nedostatek považovat za dobře splněnou. Ostatně drtivá většina odvedené práce (zejména implementační) je kvalitní, užitečná a dobře využitelná pro vědecký výzkum v oblasti koreference již v odevzdané podobě, případně pouze s jednoduchými úpravami.

Implementační část práce

	lepší	OK	horší	nevyhovuje
Kvalita návrhu *... architektura, struktury a algoritmy, použité technologie*		X		
Kvalita zpracování *... jmenné konvence, formátování, komentáře, testování*			X	
Stabilita implementace	X			

Narazil jsem na drobné problémy při instalaci/spouštění – jako shell je v Makefilech nastaveno *sh* místo *bash* (případně není nastaveno nic, takže se použije výchozí *sh*), kde ale není podporován používaný příkaz *source*, takže nástroje v odevzdané podobě nelze instalovat ani spustit. Poté, co jsem zeditoval oba Makefily, tak aby používaly *bash*, již řešení funguje výborně, v souladu s popisem v práci a bez zřejmých problémů.

Ovládání nástroje je poněkud nekomfortní, pomocí úprav konfiguračního souboru, a to nejen např. u nastavení cesty k datům (kde to má celkem smysl), ale i u volby mezi použitím českých nebo anglických dat a nástrojů (kde jde o nastavení *lang=cs* nebo *lang=en*). Pohodlnější by bylo ovládání pomocí parametrů příkazové řádky, resp. proměnných v Makefile. Alternativně by bylo vhodné, kdyby autor rovnou přiložil všechny varianty konfiguračního souboru potřebné pro vyzkoušení funkcí programu, a na příkazové řádce by si uživatel pouze zvolil, který konfigurační soubor se má použít. V Makefile by pak mohly být předpřipravené targety pro pohodlné vyzkoušení všech základních funkcí nástroje, bez nutnosti pořád dokola editovat konfigurační soubor.

Samotné zdrojové kódy jsou poměrně dobré, srozumitelné, dobře uspořádané, opatřené komentáři a kvalitním popisem v textu práce. Autor používá vhodné nástroje a knihovny – Python 3, Virtualenv, Scikit-learn, Udapi, UDPipe – tento výběr považuji za skvělý, sám bych volil stejně. Pro práci se soubory PDT bych ovšem rozhodně preferoval použití některé z Pythonových knihoven pro zpracování XML. Autor místo toho soubory zpracovává velice nízkoúrovňově pomocí rozsekávání na podřetězce a jejich porovnávání na shodu, přičemž se opírá o konkrétní formátování PDT souborů (například konkrétní způsob odřádkování), které ovšem není zaručené – závazné je v PDT souborech pouze dodržení XML schématu, spoléhání se na konkrétní formátování, které ve schématu není definováno, je nebezpečné a rozhodně jej nelze doporučit (například nemusí fungovat s novými verzemi PDT nebo se soubory zpracovanými jinými nástroji).

Celkové hodnocení Velmi dobře

Práci navrhuji na zvláštní ocenění Ne

Datum 12. června 2018 Podpis