Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	**Lenka Krippnerová**
Název práce	Identifikace složených gramatických tvarů
Rok odevzdání	2024
Studijní program	Informatika
Specializace	Umělá inteligence

Autor posudku	doc. RNDr. Markéta Lopatková, Ph.D.	**Role**	Oponent
Pracoviště	ÚFAL MFF UK		

Prosím vyplňte hodnocení křížkem u každého kritéria. Hodnocení *OK* označuje práci, která kritérium vhodným způsobem splňuje. Hodnocení *lepší* a *horší* označují splnění nad a pod rámec obvyklý pro bakalářskou práci, hodnocení *nevyhovuje* označuje práci, která by neměla být obhájena. Hodnocení v případě potřeby doplňte komentářem. Komentář prosím doplňte všude, kde je hodnocení jiné než *OK*.

K celé práci	lepší	OK	horší	nevyhovuje
Obtížnost zadání	☐	☒	☐	☐
Splnění zadání	☒	☐	☐	☐
Rozsah práce *... textová i implementační část, zohlednění náročnosti*	☒	☐	☐	☐

Komentář Prvním úkolem autorky bylo zjistit, jak jednotlivé slovanské jazyky tvoří složené slovesné tvary – k tomu bylo potřeba poměrně rozsáhlé studium tvarosloví (jde o 15 slovanských jazyků). Dále musela autorka analyzovat, jak jsou tyto tvary zachyceny v jednotlivých UD korpusech (které se mohou lišit i v rámci jednoho jazyka) – i toto je poměrně rozsáhlý úkol. Tyto znalosti pak využila pro návrh a implementaci bloků identifikujících jednotlivé složené slovesné tvary v jednotlivých UD korpusech.

Druhým úkolem bylo navrhnout obecný způsob, jak zadávat pravidla, která by – na základě anotací jednotlivých slov – identifikovala složené tvary a tuto informaci uložila do dat. Autorka zvolila pro zápis pravidel formát YAML a navrhla takovou syntax pravidel, která by měla být přístupná i pro uživatele bez programátorských dovedností; dále implementovala obecný blok, který je schopen tato pravidla načíst z konfiguračního souboru a na jejich základě doplnit požadovanou informaci do dat.

Oba tyto úkoly autorka zvládla, splnila tak zadání práce:

- navrhla, jak v rámci CoNLL-U zachytit morfologické a další rysy týkající se více slov;
- navrhla prostředí a formát pro zadávání pravidel;
- vytvořila program zpracovávající pravidla v navrženém formátu;
- otestovala vše na 15 slovanských jazycích (není jasné, zda na všech dostupných korpusech pro tyto jazyky);
- posoudila možnost nasazení na nový jazyk.

Hlavní výtkou, kterou k práci mám, je zcela chybějící evaluace alespoň na jednom jazyce (např. češtině), příp. na dalších jazycích, kde by autorka byla schopná rozpoznat správně slovesné tvary (či pro které by našla mluvčí – např. slovenština, příp. ukrajinština, ruština).

Textová část práce	lepší	OK	horší	nevyhovuje
Formální úprava *… jazyková úroveň, typografická úroveň, citace*	☐	☒	☐	☐
Struktura textu *… kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu*	☐	☒	☐	☐
Analýza	☐	☒	☐	☐
Vývojová dokumentace	☐	☒	☐	☐
Uživatelská dokumentace	☐	☒	☐	☐

Komentář Textová část splňuje běžné standardy.

- Autorka definuje dva základní cíle práce.
- Seznamuje čtenáře s UD korpusy a CONLL-U formátem. Místo plné sady rysů v UD bych přivítala seznam rysů, se kterými opravdu pracuje; pro víceslovné jednotky je UPOS je prázdný ("_"), neplatí tedy tvrzení o povinném UPOS).
- Podává souhrnnou analýzu jednotlivých slovesných tvarů. Zde lze vytknout zejm. mísení kategorie způsobu (kondicionál, imperativ (UD: rys mood), přechodník, infinitiv (UD: rys VerbForm)) a kategorie času (přítomný, minulý a budoucí, u všech je kategorie způsobu indikativ (mood)); navíc se zvláště pojímá kategorie slovesného rodu (pasivum, jen útržkovité informace platné pro indikativ), která se uplatňuje u tvarů z obou předešlých kategorií. Tyto výtky se týkají textu, nikoli navržených bloků (kde se distinkce aktivum/pasivum řeší správně v rámci bloků pro jednotlivé slovesné způsoby/časy).
- Uživatelská dokumentace poskytuje dobrý obrázek toho, jak navržené nástroje fungují a jakým způsobem se s nimi pracuje. Dále zde autorka definuje způsob psaní pravidel, vše dostatečně ilustruje na příkladech.
- Programátorský dokumentace uvádí bližší údaje o způsobu naprogramování. Poněkud překvapivě je zde i zhodnocení, jak obtížné je nasadit obecný blok pro nový jazyk a úvaha o omezeních daných zejm. nepřítomností některých informací v UD datech u některých jazyků.
- Postrádám vyhodnocení úspěšnosti jednotlivých bloků a/nebo vyhodnocení dat po aplikaci všech bloků, a to alespoň pro malý vzorek dat (viz výtku výše).

Po obsahové stránce tedy text – až na zásadní výtku týkající se evaluace a drobné výtky, z nichž jsou některé zmíněny výše – pokládám za zdařilý.

Po formální stránce je práce v zásadě bez nedostatků (pokud pomíjím drobnosti jako podivně umístěné tabulky v kap. 1, chybějící mezery před citacemi, dále kombinaci české terminologie (např. přechodník, označení časů) a terminologie vycházející z latiny (pasivum = trpný rod, imperativ = rozkazovací způsob, kondicionál = podmiň. způsob) apod.).

Implementační část práce	lepší	OK	horší	nevyhovuje
Kvalita návrhu *… architektura, struktury a algoritmy, použité technologie*	☐	☒	☐	☐
Kvalita zpracování *… jmenné konvence, formátování, komentáře, testování*	☐	☒	☐	☐
Stabilita implementace	☐	☒	☐	☐

Komentář Naprogramování jednotlivých bloků pro identifikaci složených tvarů v jednotlivých slovanských jazycích nebylo implementačně náročné, jádrem práce byla analýza problému. Práce využívá prostředí Udapi navržené pro zpracování dat UD korpusů, bylo tedy nutné se relativně podrobně s tímto prostředím seznámit.

Ani vývoj obecného prostředí, v němž je možné zadávat nová pravidla pro identifikaci dalších složených tvarů či pro zpracování nového jazyka, nebylo implementačně náročné. Hlavním přínosem je samotná metoda zápisu dat. Oceňuji implementované kontroly konfiguračního souboru, přivítala bych též kontroly unikátnost názvů (pravidel, slov, atributů v phrase, …) tak, aby se předešlo (nechtěnému) přepisování hodnot při (nezáměrné) totožnosti názvů.

Celkové hodnocení Velmi dobře (spíše lepší)
Práci navrhuji na zvláštní ocenění Ne

Datum 19. srpna 2024 **Podpis**