Posudek oponenta diplomové práce

Jméno a příjmení autora posudku:	David Mareček
Jméno a příjmení autora práce:	Jan Mašek
Název práce	Detection and Correction of Inconsistencies in the Multilingual Treebank HamleDT

Text posudku

Diplomová práce Jana Maška se zabývá detekováním a korekcí chyb v ručních anotacích závislostních korpusů.

Práce je rozdělena do šesti kapitol. Po úvodu do problematiky následuje druhá kapitola, kterou je popis souvisejících prací a postupů pro korekci morfologických značek a složkových a závislostních struktur. Zmiňuje i metodu tzv. variačních n-gramů (Dickinson and Meuers, 2003), kterou pak v experimentech používá. Třetí kapitola se zabývá kolekcí závislostních korpusů HamleDT, na kterých je detekce a korekce chyb testována. Popisuje harmonizaci korpusů do jednotného stylu, konverzi morfologických značek do Intersetu, sjednocování závislostní struktury a závislostních vztahů. Dále se zabývá nutnou úpravou formátu dat, aby byla použitelná pro existující nástroje. Čtvrtá kapitola popisuje samotnou metodu pro detekci chyb, a to pomocí variačních n-gramů (Dickinson and Meuers, 2003) pro korekci morfologických značek. Dále pak jeho úpravu pro detekci chyb v závislostech (Boyd, 2008). Pátá kapitola se zabývá experimenty a vyhodnocením výsledků. Autor použil již hotový nástroj z projektu DECCA a spustil jej na většině závislostních korpusů z HamleDT. Ukazuje statistiky nalezených variačních n-gramů, jejich pokrytí a závislost jejich relativního množství na velikosti korpusů. Dále popisuje manuální evaluaci na šesti jazycích a náhodně vybraném vzorku detekovaných, opravených a ponechaných anotací. Šestá kapitola práci uzavírá. Následuje seznam použité literatury a seznam zkratek. v příloze jsou pak statistiky jazyků v HamleDT a popis datových formátů. K práci je přiloženo DVD se všemi nástroji a skripty použitými v experimentech.

Diplomová práce má celkem 67 stránek, z toho vlastní text práce tvoří 45 stránek. Práce je přehledně strukturovaná a je psána dobře srozumitelnou angličtinou. Použitá literatura je řádně citována. Na některých místech je text příliš stručný. Například v popisu algoritmu a heuristik pro detekci chyb v závislostní struktuře bych uvítal lepší znázornění toho, co přesně obsahuje "variation nuclei" a jaký kontext se uvažuje v heuristikách. Jde například pouze o slovní formy nebo i o morfologické tagy?

Práce obsahuje v zásadě dvě témata: jedním je harmonizace závislostních funkcí a druhým je samotná detekce a korekce chyb. V diskusi následující za experimenty autor zmiňuje hlavně problémy

s harmonizací závislostních funkcí z důvodu nekonzistence harmonizačních bloků a chybějících anotačních manuálů. Čekal jsem zde spíše příklady toho, co se opravilo správně a co špatně a jaký vliv na to měly vnitřní anotační nekonzistence v rámci jednoho korpusu, nikoli nekonzistence mezi korpusy.

Student splnil zadaný úkol. Použil již existující nástroje pro detekci chyb z projektu DECCA a stačilo tedy relativně málo implementační práce spočívající hlavně ve psaní konverzních a vyhodnocovacích skriptů. Na druhou stranu dokázal vést tým několika anotátorů a zpracovat kvalitní evaluaci výsledků včetně očekávaného vylepšení závislostního korpusu, které se pohybuje od 0 do 0.24% tokenů.

Otázky:

1. Pokud se v rámci korekce převěšuje závislostní hrana, jakým způsobem řešíte možnost vytvoření cyklu?

2. Proč je nutná pro správnou detekci a korekci chyb normalizace závislostních funkcí napříč korpusy? Přijde mi, že experimenty na jednotlivých jazycích běží nezávisle na sobě.

3. Trénoval jste závislostní parser na datech s již opravenými morfologickými značkami, nebo na původních datech? Dovedu si představit, že chyby v anotaci závislostí mohou často vznikat tím, že anotátor se řídí značkou, která je špatně.

4. Zkoušel někdo evaluovat závislostní korpus vzniklý pouhým přetaggováním (přeparsováním), bez předchozí detekce chyb? Toto by nadělalo pravděpodoně víc chyb, než oprav, ale bylo by zajímavé to porovnat s Vaším přístupem, aby bylo vidět, kolik zbývá prostoru pro další vylepšování anotací.

Doporučení k obhajobě

Z výše uvedených důvodů práci *doporučuji* k obhajobě.

Soutěž studentských prací

Vynikající práce vhodná soutěže studentských prací: **NE**.

V Praze dne 29. 5. 2015

Podpis: