Posudek vedoucího diplomové práce

Jméno a příjmení autora posudku: Zdeněk Žabokrtský

Jméno a příjmení autora práce: Jan Mašek

Název práce: Detection and Correction of Inconsistencies in the Multilingual Treebank HamleDT

Vlastní text:

Popis práce

Předložená práce je zaměřena na automatické rozpoznávání a opravy chyb v syntakticky anotovaných korpusech, které jsou součástí kolekce HamleDT. Vyvinuté postupy jsou v rámci možností jazykově nezávislé a jednotným způsobem zpracovávají morfologické a závislostní značkování. Jádrem řešení detekce nekonzistencí je přejatá a upravená metoda variačních n-gramů. Každý variační n-gram obsahuje variační nukleus a jeho kontext, přičemž hlavním východiskem metody je předpoklad, že odlišně značkovaný totožný nukleus v totožném kontextu může signalizovat chybu. Opravy nalezených korpusových pozic jsou pak založeny na použití existujícího morfologického taggeru a závislostního parseru. Detekované i automaticky opravené chyby jsou vyhodnoceny na ručně anotovaném vzorku pro několik jazyků. Z vyhodnocení experimentů plyne, že vyvinutý postup lze použít k redukci chyb v korpusu HamleDT.

Včetně seznamu literatury a příloh má práce 76 stran. Práce je psána anglicky. Součástí práce je DVD s vyvinutými zdrojovými kódy, použitým softwarem třetích stran a extrahovanými daty.

Hodnocení

Kladně hodnotím zejména následující aspekty předložené práce:

- Autor se podrobně seznámil s řadou souvisejících zahraničních publikací a s mnoha jednotlivými technickými i lingvistickými problémy dat zahrnutých do kolekce HamleDT.
- Autor byl schopen do řešení své úlohy zapojit několik různých existujících softwarových nástrojů pro zpracování přirozeného jazyka.
- Autor zorganizoval ruční anotaci pro několik typologicky odlišných jazyků, své experimenty pečlivě vyhodnotil a prokázal, že navržený postup vede k pozitivním výsledkům a je robustní.
- Práce je přehledně strukturovaná a psaná výbornou angličtinou, ve finální verzi práce jsem nalezl jen zcela minimální množství jazykových chyb.

Pokud jde o slabší místa předloženého textu, všechny mé kritické poznámky mají společného jmenovatele: je škoda, že se práci nepodařilo dotáhnout ještě o pár kroků dál. Předem bylo zjevné, že jazyková i datová rozmanitost kolekce HamleDT povede k mnoha neočekávaným problémům, nicméně podle mého názoru se u několika podúloh a formulovaných hypotéz autor zastavil na první technické překážce (např. dvakrát v sekci 5.1.1, dvakrát v sekci 5.1.2, v sekci 5.2.2, dvakrát v sekci 5.3.2). Z textu také není zřejmé, zda a jak konkrétně se vyvinutý postup projeví na budoucích verzích kolekce HamleDT.

Doporučení k obhajobě:

Z výše uvedených důvodů práci *doporučuji* k obhajobě.

Vynikající práce vhodná pro soutěž studentských prací	ANO ☐

Seznam soutěží studentských prací, viz http://www.mff.cuni.cz/studium/bcmgr/prace/

Pokud jste výše zaškrtli ANO, zdůvodněte prosím svůj návrh, případně uveďte konkrétní soutěž, pro kterou je práce vhodná (rámeček lze nechat prázdný, pokud za dostatečné zdůvodnění považujete text posudku):

V Praze dne: 27. 5. 2015

Podpis:**

** nehodící se škrtněte (vymažte)*

*** do SISu vkládejte formulář nepodepsaný (ve formátu PDF), podpis je potřeba doplnit až na vytištěný posudek.*