Posudek *vedoucího* diplomové práce

Jméno a příjmení autora posudku: Markéta Lopatková

Jméno a příjmení autora práce: **Adéla Limburská**

Název práce: **Semantic information from FrameNet and the possibility of its transfer to Czech data / Sémantická informace ze sítě FrameNet a možnosti jejího využití pro česká data**

Vlastní text (sem prosím napište text posudku, délka textu posudku není omezena):

Diplomantka se ve své práci zabývá možnosti využití sémantické informace ze sítě FrameNet a jejím přenosem do českých jazykových zdrojů. Pracuje s FrameNetem, odkud čerpá sémantické informace, dále s paralelním korpusem PCEDT, s nástroji pro automatický překlad dat (eng->cz) a s knihovnami pro strojové učení (AI::DecisionTree Module [Williams,2012] pro rozhodovací stromy a LIBSVN [Chang and Li,2011] pro SVM).

Po úvodní kapitole seznamující s daty FrameNetu se diplomantka věnuje vzorku plně anotovaných dat z PropBanku (Full Text Annotation), jeho převodu pomocí PCEDT a evaluaci získaných dat (celkem 337 vět s plnou anotací, 418 různých sémantických rámců (dále SF), 2 121 instancí SF, úspěšnost 79% pro přiřazení SF). Vzhledem k poměrné malému vzorku vět se její další práce soustředila na rozšíření dat pomocí překladu příkladových vět u jednotlivých lexikálních jednotek (LU); z automaticky přeložených dat (poskytnutých R. Rosou) tak získala 170 tis. párů anotací (pro 10 tis. LU, s úspěšností 71% pro přiřazení SF na vzorku 176 párů vět).

Diplomantka srozumitelně seznamuje čtenáře s anotačními schématy pro jednotlivé zdroje, oceňuji především podrobnou analýzu dat a analýzu jejich kvality – Adéla podává kvantitativní údaje, na vzorcích dat provádí i lingvistické vyhodnocení chybovosti a zamýšlí se nad zdroji chyb. Nezanedbatelné úsilí věnovala též automatickému i manuálnímu čištění dat. Tuto část práce tedy hodnotím velice pozitivně.

Druhým úkolem byl návrh systému pro automatickou predikci SF. Diplomantka využila, příp. mírně upravila existující nástroje, které trénovala pro 100 LU s nejvyšší frekvencí v trénovacích datech, omezila se přitom na základní experimenty. Podává popis rysů pro jednotlivé experimenty, u rozhodovacích stromů se též zamýšlí nad jejich lingvistickou motivací a zdůvodněním výsledků. Protože výsledky experimentů nedosahovaly předpokládané kvality, vrátila se Adéla k analýze a další manuální revizi a čištění dat. Tím dosáhla úspěšnosti přiřazení SF kolem 76%, resp. 68% při počítání jen na slovech evokujících SF (ale ovšem i vysoké baseline počítané jako úspěšnost při přiřazení nejčastějšího SF). Vzhledem ke kvalitě trénovacích dat (viz výše) zůstává otázkou, jaká je (teoretická) mez predikce SF pomocí metod strojového učení. Dále je nutno konstatovat, že přenos sémantické informace mezi jazyky je obtížnou problematiku, kvalitní řešení přesahuje možnosti běžné diplomové práce.

K řešení diplomové práce mám jedinou zásadnější výhradu – diplomantka se zcela soustředila na úlohu přiřazování SF, rezignovala na úlohu identifikace jednotlivých elementů SF, již zadání práce též předpokládalo. Tuto informaci nezkusila využít např. ani pro rysy při přiřazování SF (je samozřejmě otázkou, zda by k této úloze měla dostatečná či dostatečně kvalitní data).

Práce je psaná dobře čitelnou angličtinou, našla jsem minimum chyb či překlepů (např. kap. Introduction: FrameNet obsahuje popis cca 10 tisíc lex. jednotek a 1 tisíc SF, nikoli 1 tis. lex. jednotek).

Text práce má standardní strukturu, k textu je přiloženo CD obsahující jazyková data (či jejich vzorky), vlastní skripty diplomantky, použité nástroje pro strojové učení a soubor README se stručným popisem dat a skriptů.

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: 25.8.2016

Podpis: