

Posudek oponenta diplomové práce

Autor práce: Marek Blahuš

Název práce: Extending Czech WordNet Using a Bilingual Dictionary

Jméno oponenta: Zuzana Nevěrilová

Diplomová práce Marka Blahuše *Extending Czech WordNet Using a Bilingual Dictionary* (Rozšiřování českého WordNetu pomocí překladového slovníku) se zabývá převodem anglicko-českého slovníku do elektronické podoby s obsahovými značkami, nalezením vhodného algoritmu pro přidávání nových synonymických řad (synsetů) do českého WordNetu a implementací takového algoritmu. Práce se skládá z písemné a praktické části.

Cílem práce bylo rozšířit existující lexikální síť WordNet o nové synsety. Zdrojem pro rozšíření byl Velký anglicko-český slovník vydavatelství LEDA (dále VAC).

Písemná část je rozdělena do šesti kapitol: v úvodu autor popisuje motivaci a stávající data. Ve 2. kapitole se autor věnuje podobným snahám pro jiné jazykové páry. Z analýzy těchto projektů vychází při návrhu vlastního postupu. Druhá kapitola rovněž obsahuje přehled použitého software. Ve třetí kapitole se zabývá analýzou VAC a převodem na obsahové značkování. Autor vycházel jednak z vysvětlivek ve slovníku, jednak z vlastních pozorování. Ve čtvrté kapitole, která je oproti předchozím poměrně krátká, autor popisuje generování nových synsetů a rozšíření českého WordNetu. Pátá kapitola obsahuje číselné vyjádření objemu práce. Poslední, závěrečná kapitola obsahuje výhledy pokračování práce do budoucna.

Práci je možné rozdělit na dvě části: velmi pracné převedení prezentačního značkování elektronické verze VAC na obsahové. Autor zjevně věnoval tomuto kroku značné úsilí, protože výsledná data jsou téměř stejného rozsahu a zároveň kvalita výstupu je dobrá.

Druhou částí bylo vyhledávání anglicko-českých překladových párů. Autor, inspirován podobnými pracemi pro jiné jazykové páry, navrhl vlastní algoritmus vyhledávání párů a napojení nových synsetů na existující synsety ve stávajícím českém WordNetu. Právě vyhledávání vhodných hyperonym pro nové synsety se zdá být největším úskalím práce. Například VAC uvádí u výrazu „stray“ tyto překlady: [animal] zatoulané zvíře; [dog] zaběhlý or zatoulaný pes; [sheep] bludná ovce; [bullet] zbloudilá kulka or (sic!) střela. V anglickém WordNetu je **stray:1 (n)** hyponymem od **domestic cat:1**. V českém WordNetu potom vznikne hyponymum od „kočka domácí“, které obsahuje literály „bludná ovce“ či „zbloudilá střela“. Bylo by možné při rozhodování o hyperonymu vzít v úvahu informace uvedené v prvku IEN (zde v hranatých závorkách)? Je možné odhadnout, ve kterých případech špatné přiřazení do hypo/hyperonymického stromu nastává? Kolik je takových případů?

Práce má řadu užitečných vedlejších efektů. Nejdůležitějším je bezesporu získání rozsáhlého a dobře značkováného překladového slovníku. Dalším je zpětná vazba pro morfologický analyzátor **ajka**, vyvíjený na FI MU. Autor použil extrahované literály jako vstup pro morfologický analyzátor, čímž zjistil slova, která morfologickému analyzátoru pravděpodobně chybí. Velkým kladem práce je rozdělení skriptů na méně obsáhlé úlohy, takže je možné je používat zvlášť a extrahovat jejich pomocí různé informace. Skripty tedy nejsou „na jedno použití“.

Navržené otázky k diskuzi: jaký výsledek by uvedená metoda měla na existující data českého WordNetu? V širším kontextu: jsou data ručně zpracovaná experty (dosavadní WordNet) srovnatelné kvality s daty pořízenými veřejností (Wiktionary, Wikipedie) či daty pořízenými automaticky (např. metodami popsány v této práci)?

Příložené CD obsahuje zdrojové kódy software a dokumentaci. Drobnou výhradou je nedodržení konvence při pojmenování třídy `transform.java`. Práce je psána dobrou a většinou srozumitelnou angličtinou, jen s občas se vyskytujícími překlepy (např. „for instace“ na str. 11 či „buy specifying arguments“ na str. 26). Podle mého názoru by textu popisujícímu procedury (např. v sekci 3.3) prospěl symbolický zápis (autor by se tak vyhnul méně obratným slovním spojením jako „Czech literal“, „another Czech literal“, „the Czech literal in question“). Mám výhrady k seznamu literatury, kde *Slovník spisovné češtiny* nemá uvedeného ani jednoho autora. U elektronických zdrojů chybí datum získání zdroje. Práce je dobře logicky členěna a doplněna přílohou obsahující příklady nově získaných synsetů.

Práci doporučuji k obhajobě a navrhuji známku A až B podle výsledku obhajoby.

19. června 2011
Zuzana Nevěřilová