Oponentský posudek doktorské disertační práce

Václav Klimeš: Analytical and Tectogrammatical Analysis of a Natural Language

Disertační práce Václava Klimeše se zabývá automatickou syntaktickou analýzou přirozeného jazyka, zejména češtiny. Ve zmíněné oblasti má práce široký záběr. Věnuje se jak analytické rovině, tak i rovině tektogramatické (s praktickými omezeními a bez koreference a bez aktuálního členění), věnuje se jak budování stromových struktur, tak i jejich značkování analytickými resp. tektogramatickými značkami. Disertační práce využívá Pražského závislostního korpusu jako základního lingvistického materiálu a na těchto datech zkouší existující, případně modifikuje či buduje vlastní algoritmy, pomocí kterých automaticky, tedy ne pomocí ručně psaných lingvistických pravidel, získává potřebnou znalost gramatiky češtiny. Předvedené algoritmy jsou obecné, tj. ne nutně aplikovatelné pouze na češtinu.

Disertační práce Václava Klimeše obsahuje čtyři kapitoly a její struktura je dále obohacena o stručný index základních termínů, přehled tabulek a obrázků, bibliografii a tabulkový přehled hodnot vybraných atributů včetně popisu části redukovaných a v práci použitých morfologických značek (příloha). První kapitola je úvodní a čtvrtá kapitola obsahuje stručný závěr. Jádro práce je obsažené v kapitole druhé a třetí. Druhá kapitola je věnovaná analytickým stromům a třetí kapitola stromům tektogramatickým. Práce je psaná anglicky.

Z hlediska využití automatických metod učení, které tvoří základ práce, autor v druhé kapitole uvádí dvě metody: metodu řídící se délkou kontextu (2.3 Method 1: Optimal Context Length) a metodu klasifikace řízenou pravidly (2.4 Method 2: Transformation-based classification). V obou případech je základem Brillův přístup řízený pravidly, který autor modifikuje. Uvedené modifikace jsou obsáhlé a s dostatečnou originalitou. Z hlediska dosavadního výzkumu týkající se metod řízených pravidly disertační práce Václava Klimeše přináší důležitý dodatek, tj. že strukturní analýza syntaktického typu je přístupu řízeném pravidly lépe určena, pokud se tento přístup uchopí jako přístup klasifikační (oproti pravidlům provádějící transformace přímo nad stromovou strukturou, jak tomu bylo doposud, a to jak pro složkový tak i pro závislostní parsing).

Dalším přínosem této práce je celková šíře záběru i z hlediska implementace: vstupní morfologicky označkovaná věta je pomocí analytického rozboru (vč. analytických funkcí) obohacena o povrchovou strukturu v podobě analytického stromu, a dále je tento analytický strom převeden dalším krokem na strom tektogramatický (vč. přidání funktorů). Obvykle se dosavadní práce věnují buď povrchové, či hloubkové syntaxe, a to z hlediska získávání stromových struktur, nebo z hlediska anotace stromových struktur pomocí značek (analytické či tektogramatické). Relativně nevelkou úspěšnost u analytických stromů vynahrazuje relativně vysoká úspěšnost u části tektogramatických stromů vč. funktorů.

Do třetice, autorovým přínosem je i rozpracování problematiky porovnání tektogramatických stromů, kterou popisuje v odd. 3.2 a dále i implementuje.

V následující části bych chtěl autora upozornit na problémy, jejichž vyřešení/odstranění by mohlo úroveň práce zvýšit.

1. Terminologie: je použita řada termínů a jejich uvádění do textu je implicitní a na místech nedostatečně přesné. Vzhledem k implementačnímu charakteru práce a vzhledem k šíři záběru bych uvítal strukturovanější styl a přehlednější vymezení pojmů převzatých a zejména pak pojmů vlastních.

2. Neurčitost: často se v práci uvádí slova typu, "usually called" v případech, kdy jasná definice existuje (např. 3. poznámka pod čarou, kde takto může vzniknout domněnka, že rozdělení do více souborů je stand-off anotace), "somehow" v případech, kdy by bylo v zájmu práce vysvětlit jak (např. odd. 2.3.6, první věta), "certain attribute" kdy by bylo lépe specifikovat, "enourmous number of factors", aniž by dimenze problému byla blíže určena popisem nebo odkazem, atd. I když zde uvádím jen jednotlivé příklady, je možné uvést řadu dalších příkladů obdobné autorové neopatrnosti.

3. Vysvětlení/reference: jsou místa, která by si zasloužila podle mého názoru větší pozornost a komentovanou referenci. Týká se to zejména:
   - úvodní části tektogramatické analýzy jak z hlediska lingvisticky teoretického tak z hlediska anotačního (např. v odd. 3.1 "Annotation at the t-layer is a very complex task ...." je uvedeno bez dalších vysvětlení či odkazů);
   - využití dvoupoziční redukce morfologických značek či uvedení jiných redukcí (odd. 2.4.1);
   - evaluační postupy jiných autorů v případě evaluace tektogramatických stromů tak, aby šlo lépe porovnávat získané výsledky (odd. 3.3).

4. Uvítal bych jasnější oddělení vlastního přínosu od přínosu dosavadní práce ostatních. Dále by bylo záhodno zvýšit čitelnost zejména formálnějších části textu; typ takového příkladu je na str. 53, "In the case of a parent, ... ".

Připomínky uvádím spíše jako konstruktivní návrhy pro autora. Svůj posudek uzavřel tím, že máme před sebou práci se širokým záběrem, máme před sebou práci, v které autor samostatně řeší komplexní otázky syntaktického rozboru a máme před sebou též sadu nástrojů, které jsou vzájemné propojené, a které lze využít i s relativně malým počtem trénovacích dat.

Práci doporučuji k obhajobě.

Praha, 23. srpna 2006