Vyjádření vedoucího doktorské disertační práce

Václav Klimeš: Analytical and Tectogrammatical Analysis of a Natural Language

Předložená práce popisuje způsoby povrchově a zejména hloubkově syntaktického parsingu přirozeného jazyka a aplikuje tento postup na češtinu, na které dané postupy vyhodnocuje. Práce je členěna na čtyři oddíly (kapitoly), z nichž druhý a třetí oddíl (analytická a tektogramatická analýza) popisuje vlastní práci autora. Práce je doplněna mnoha obrázky a tabulkami, které jednak vysvětlují problematiku práce a jednak ilustrují popsané metody a algoritmy; na závěr práce je strukturovaný seznam tzv. funktorů (sémantických značek pro typy závislostních vztahů), jejichž přiřazovaní patřilo do škály problémů v práci řešených. Značná pozornost je věnována i vyhodnocování úspěšnosti u tektogramatického parsingu, neboť tento problém je na rozdíl od parsingu analytického (povrchového) značně netriviální.

Autorovým cílem bylo najít obecné metody pro parsingu, a to zejména pro parsingu hloubkový (tektogramatický). Speciálním požadavkem bylo, aby autor volil takové metody, které by byly použitelné i pro jiné jazyky (za předpokladu stejné nebo velmi podobné specifikace reprezentace analýzy věty na tektogramatické rovině). Pro statisticky založenou analýzu autor zvolil tzv. transformační metodu popsanou poprvé Brillem v r. 1995 a efektivně implementovanou v r. 2001 Florianem a Ngai.

Výsledky autorem navržených metod parsingu na rovině povrchové syntaxe (analytické) jsou pro přiřazení struktury horší než současné nejlepší metody parsingu udávají (dependency accuracy 74.6%, tj. horší o cca 8 procentních bodů [parser McDonalda et al.]); vyzkoušel však dvě nové metody v dosud jinak početných pracích nepoužité, a prokázal tak, že zřejmě ani značné další úsilí (vzhledem k dosaženému rozdílu ve výsledcích) k průkazně lepším výsledkům nepovede. Jeho výsledky přiřazování tzv. analytických funkcí však předčily dosud nejlepší výsledky (byť nepublikované, autor: Z. Žabokrtský), a v absolutních číslech jsou velmi uspokojivé (94.3% i při plně automatickém předzpracování taggerem a parserem).

Na tektogramatické rovině je práce autora průkopnická a prakticky první, která pojímá danou problematiku komplexně; jeho výsledky tedy nastavují pomyslnou lať ku, kterou všechny následné práce musí překonat. K dispozici měl pouze tzv. „automatickou proceduru" předzpracování tektogramatických stromů pro anotaci A. Böhmové, svoje vlastní předchozí experimenty a pravidla napsaná Z. Žabokrtským (specificky pro angličtinu, a přiřazování tektogramatických funktorů pro češtinu). Tektogramatická rovina je velmi komplexní; počítáme-li i technické atributy uzlů, dojdeme k číslu 39 (vs. 5 na rovině analytické). Autor se zaměřil na strukturu tektogramatického stromu a s ní bezprostředně související atributy – a to celkem na 13 z nich, které jsou klíčové pro strukturální anotaci. Šest z nich zpracovával pomocí obecných statistických metod, zbytek obecnými jazykově v zásadě nezávislými pravidly. Autor správně konstatuje, že zbývající atributy (s výjimkou koreferenčních a atributu t/f aktuálního členění) je možno s vysokou přesností určit manuálně napsanými pravidly, alespoň za současného stavu anotačních pravidel. Autor pracoval s češtinou, neboť dosud existují (trénovací a testovací) data prakticky pouze pro ni (korpus PDT 2.0).

Zpracovávaných šest strukturálních atributů je však z celé tektogramatické roviny nejobtížnější. Jedná se o strukturu stromu, funktor, tektogramatické lema, typ uzlu, sémantický slovní druh, a význam slovesa (resp. jeho valenční rámec). Jak již bylo řečeno, tyto výsledky lze s předchozími pracemi srovnat jen obtížně, neboť takové komplexní práce neexistují; výsledky v některých jednotlivých bodech jsou však nepochybně nesrovnatelně lepší (struktura, funktory).

Za největší a velmi podstatný přínos autora považuji:

- podrobnou analýzu možností tektogramatické analýzy, a její konceptuální i technologické rozdělení na 4 fáze, které je originální a nepochybně zůstane východiskem i pro navazující práce, i kdyby používaly jiné metody analýzy;
- matematickou a algoritmickou specifikaci evaluace výsledků tektogramatické analýzy, která nastavuje standard, využívaný v budoucnu všemi následníky autora jistě po velmi dlouhou dobu;
- dotažení strukturální tektogramatické analýzy do aplikačně použitelného stavu (výše uvedených 13 atributů);

- vypracování metody přiřazování analytických funkcí, která je dosud nejlepší možnou;
- podrobnou analýzu chyb navržených a implementovaných metod, a to jak na analytické, tak tektogramatické rovině.

Pokud lze práci něco vytknout, nejedná se o to, že na analytické rovině nebylo dosaženo srovnatelných světových výsledků (autor podrobně analyzoval chyby jeho dvou navržených řešení a ušetřil tak práci s touto analýzou pro budoucí pokusy o vylepšení analytického parsingu), nýbrž na některé pasáže zdůvodňující navržený postup, které nejsou zcela jasné: (a) kap. 2.6.3, první odstavec říká, že není možno užít v podmínkách pravidel již přiřazené s-tagy; metoda transformačního učení však obecně takové podmínky připouští (je třeba lépe vysvětlit, proč v daném případě toto není možné); a (b) kap. 2.3.5 – zde chybí zmínka o tom, proč stanovení prahu vyloučení „nespolehlivých pravidel" nebylo doladěno také na trénovacích datech, resp. jejich oddělené části.

Práce je psána anglicky, jazyk je srozumitelný, byť ne vždy formulačně (jazykově) obratný; při případném publikování některých částí práce doporučuji hlubší jazykovou korekturu. Ani po formální stránce k ní nemám výhrady.

Závěr: celkově práci považuji za velmi dobrou a doporučuji, aby byla přijata a obhájena jako práce disertační.