Posudek disertační práce (Alexandr Rosen)

Mgr. Jiří Mírovský
Netgraph – A Tool for Searching in the Prague Dependency Treebank 2.0

Autor vytvořil vyhledávací jazyk a odpovídající softwarový nástroj pro hledání v syntakticky anotovaném korpusu a pro zobrazování nalezených dat. I když obojí šil na míru víceúrovňovému Pražskému závislostnímu korpusu (PZK), výsledek lze použít i pro jiná data, jak ukazuje příloha E.

Autor navázal na předchozí úsilí v tomto směru, ale zároveň vyšel z požadavků, co vše mají umět vyjádřit dotazy do velmi bohatě anotovaného PZK (kapitola 2). K možným dotazům na lingvisticky složité jevy a konstrukce našel odpovídající formální vyjádření v dotazovacím jazyce, a to v grafickém a textovém formátu (kapitola 3), a na mnoha příkladech ukázal, že tento jazyk splňuje výchozí požadavky (kapitola 5).

Završením autorovy práce je vyhledávací nástroj Netgraph (kapitola 6). Přiložené CD obsahuje soubory pro instalaci klienta a serveru, dokumentaci, ukázky dat a četné autorovy publikace (konferenční příspěvky), vztahující se k Netgraphu.

Dotazovací jazyk je podroben důslednému srovnání s jinými prostředky pro hledání v syntakticky anotovaných korpusech. Ze srovnání vychází velmi dobře (kapitola 6). Další test představuje analýza interakcí uživatelů veřejně přístupného serveru Netgraph, kteří podle všeho plně využívají možnosti nástroje (kapitola 8).

Text práce je přes drobné formální a jazykové nedostatky dostatečně srozumitelný a přehledný, včetně příkladů a příloh. K tomu přispívá i vysoká úroveň grafického zpracování. Software na přiloženém CD musí překonat očekávání každého, kdo s Netgraphem dosud neměl tu čest.

Následující výhrady a připomínky v žádném případě nezpochybňují význam a kvalitu práce, o kterých se zmíním v závěru.

Skryté uzly

Syntaktická anotace na rovině analytické a tektogramatické se často liší nejen hodnotami atributů, ale i stukturně, počtem hran a uzlů a směrem závislosti mezi nimi. Proto by bylo vhodné mít možnost paralelního zobrazení na více rovinách, třeba tak, jak je to na obrázku na str. 28. Data v PDT 2.0 však Netgraphem takto zobrazovat nejdou. Na str. 21 autor uvádí dvě základní možnosti přístupu: přímý přístup k analytické rovině, bez možnosti zobrazení roviny tektogramatické, nebo přímý přístup k tektogramatické rovině, s přístupem k analytické rovině „skrze" rovinu tektogramatickou. Jedině v tomto druhém případě lze zobrazit informace z obou rovin, ale nikoli paralelně, nýbrž s využitím pomocných („skrytých") uzlů. Ty rozšiřují tektogramatický strom o části, které často na žádné z rovin nejsou, a z nichž bez znalosti zásad anotace nelze podobu analytického stromu odvodit. Není tedy pravda, že *we either see/search/study the analytical layer with all information from the lower layers available, or the tectogrammatical layer,* ***also with all the information from the lower layers available***. Povrchově syntaktickou strukturu nelze zobrazit, pouze (poměrně náročným

způsobem) vyhledávat. Zdá se, že není ani možné si nějakým jednoduchým způsobem k nalezenému tektogramatickému stromu nechat vyhledat odpovídající strom analytický.

U zvoleného řešení pro hledání a zobrazování výsledků na obou „vysokých" rovinách vidím tyto nevýhody:
- analytický strom je nepřístupný, nebo jen částečně a s obtížemi, oklikami, či protiintuitivně (jako třeba u dotazu na obrácené závislosti na str. 72)
- tektogramatický strom se skrytými uzly zobrazuje dvě odlišné informace velmi podobným způsobem a může být proto obtížně čitelný
- skryté uzly komplikují použití metaatributů, které se liší v tom, zda je berou nebo neberou v úvahu

V práci jsem nenašel přesvědčivé argumenty pro zvolené řešení a proti paralelnímu zobrazení obou/všech rovin. Prosím o ně při obhajobě.

Méně podstatné věci, náměty na rozšíření, poznámky:

Pokud autor v části 2.1.2 uvádí mezi vyhledávacími nástroji i Manatee/Bonito, který neumí hledat v treebanku, měl by se zmínit i o IMS Corpus Workbench/CQP. Dalším kandidátem na zmínku by mohl být Poliqarp, který lze použít i pro hledání v syntakticky značkovaných korpusech (http://korpus.pl/~adamp/Papers/2007-acl-demo/acl07.pdf).

V části 2.3.1 by bylo vhodné uvést příklady u všech jevů, oponentovi chybí zejména u složených přísudků, kvazi-fokusu a neprojektivit.

Str. 29, část o shodě: podmět s přísudkem se v češtině shoduje i v osobě. Predikátová adjektiva se navíc shodují i v pádě (pokud nejsou v instrumentálu): *Jen pět bodů bylo nejasných.*

Zdá se mi, že uživatel by jako základní způsob hledání každé závislosti neměl zadávat závislost pomocí hran, ale využít atributu eparents. To ovšem souvisí se způsobem značkování PDT, přesto bych chtěl autora poprosit o extrémně hypotetickou úvahu na téma využití „efektivních" závislostí jako výchozích při zadávání dotazů.

Termíny se někdy objevují poprvé bez vysvětlení, např. už v úvodní části 1.4 o PDT se předpokládá, že čtenář ví, co jsou gramatémy, hloubkový slovosled, textová a gramatická koreference. Tato část by vůbec chtěla rozšířit o pasáže objasňující lingvistickou motivaci jednotlivých rovin, možná vůbec není nutné ji dávat do úvodu.

V části 2.3.2 se na str. 27 uvádí příklad, kdy je třeba vyhledávat na dvou rovinách současně, ale následující obrázek výše uvedenému konkrétnímu příkladu neodpovídá.

Příklad na str. 65 nahoře vypadá jako příklad na kvazi-fokus. Je to tak? Pokud ano, mělo by to tam být uvedeno.

Str. 69 dole, pasáž o (ne)projektivitě: nezajímavost tektogramatické roviny pro zkoumání projektivity patrně plyne z teoreticky zdůvodněné preference projektivních konstrukcí na této rovině. Není-li tektogramatická rovina pro studium neprojektivit dost relevantní, mohla by to být indicie svědčící pro revizi příslušných teoretických předpokladů.

11.5, příloha E (str. 134 a dále): Bylo by hezké mít u dalších korpusů nějaké údaje o jejich velikosti, našel jsem je jen u latinského treebanku.

Připomínka k Netgraphu jako vyhledávači: atributů i jejich hodnot je moc. Možná by to chtělo uživateli pomoci s orientací, třeba rozdělením na více kategorií a/nebo nějakými bublinami.

Drobnosti – angličtina, terminologie, překlepy:

- str. 13, 9. řádek od konce: má být *thesis* místo *theses*
- str. 12 a 13: chybné uzavírací uvozovky
- str. 12 a 13 i dále: chybné užívání spojovníků místo pomlček
- str. 14: *function words* lepší než *functional words*
- str. 18, poslední řádek: má být *trees whose head* místo *trees, whose head* (jde o těsné/restriktivní rozvití)
- str. 22, 7. řádek od konce: má být *most complete* místo *the most complete*
- str. 33, 2. řádek: lépe *absence* místo *non-presence*
- str. 37: lépe *positive integer* místo *positive whole number*
- str. 53, bod o atributu `eparents`: lépe asi *an identifier of each linguistically effective father for each node*, o tomto pojmu se tu mluví poprvé, jestli jsem něco nepřehlédl, chtělo by to vysvětlení, třeba už v souvislosti s koordinací na str. 23
- str. 141, pozn. pod čarou: *clickingg*

Závěr

Disertace Mgr. Mírovského je úctyhodné dílo. Podařilo se mu vyřešit problém přístupu ke složitě strukturovaným lingvistickým datům tak, aby sloužily uživatelům bez nutnosti dlouhého školení nebo studia formálních konstruktů dotazovacího jazyka. Grafické vyjádření dotazu i postup jeho zadávání je dostatečně intuitivní. Velmi důležitá je i možnost zadat dotaz textově nebo porovnávat obě podoby. O užitečnosti autorovy práce svědčí statistika využití systému i jeho nasazení na korpusy v jiných jazycích. O její vědecké úrovni a originalitě řešení není pochyb, navíc autor o svém vědeckém přínosu přesvědčil na důležitých mezinárodních konferencích. Disertace nepochybně prokazuje předpoklady autora k samostatné tvořivé práci. Jsem přesvědčen, že titul Ph.D. mu právem náleží.

[redacted]

26. července 2008