

Oponentský posudek disertační práce

Název: *Automatic Analysis of Temporal Relations within a Discourse and Its Application within a Machine Translation Framework*
Autor: Petr Němec
Obor: matematická lingvistika, MFF UK
Rozsah: 85 stran textu
4 textové přílohy
1 DVD

Předložená práce má jako hlavní sjednoudující téma analýzu časových vztahů v diskurzu. V jednotlivých kapitolách autor popisuje v podstatě izolované projekty spojené uvedeným tématem – anotaci časových vztahů, automatickou analýzu těchto vztahů v anotovaném korpusu, systém generování textů ze závislostní struktury věty, formalizmus pro vyhledávání a nahrazování v korpusu stromů a pokus o vylepšení generování textů pomocí analýzy časových vztahů.

Text práce je napsán anglicky na slušné jazykové úrovni, obsahuje ovšem několik zbytečných překlepů (i typu *does* místo *does not* či překlepy ve vzorcích). Sazbě textu se dá vytknout zejména uvádění malých písmen v položkách seznamu literatury i na místech zkratk a označení národních jazyků (nlp, czech, english, ...). Autor také v textu často chybně zaměňuje znaky ' ' a ' '.

Přednosti a přínos práce:

- a) vytvoření konkrétního anotačního schématu pro popis časových vztahů a jeho použití při označování referenčního korpusu
- b) implementace pravidlového systému pro analýzu časových vztahů převážně podle dříve známého principu
- c) implementace systému na generování vět
- d) návrh a implementace nového formalizmu pro vyhledávání a nahrazování v korpusu závislostních stromů

Nedostatky a problematické body v práci:

- koncepční nedostatky:
 - a) podle tématu práce by bylo více než vhodné věnovat část textu obecnějšímu úvodu do problematiky, který by zcela jistě měl obsahovat např. odkazy na temporální logiky a logické systémy, které se zabývají časovými vztahy z hlediska logické analýzy jazykového významu (nabízejí se zde práce z oblasti transparentní intenzionální logiky – Tichý, Materna, Koukolíková, ...).
 - b) práce je těsně vázána na podkladovou teorii spjatou s PDT, autor nenabízí dostatečné srovnání s jinými přístupy.

- c) z textu není zřejmé, proč se místo vlastního návrhu anotačního schématu nepoužilo zmiňované schéma TimeML.
 - d) str. 37-38, sekce 3.2 a 3.3 – popis *Content Clauses* je podle mého názoru značně nepřehledný a popis *Recursive Temporal Principle* pokládám za nedostatečný a nedůsledný. Autor v sekci 3.3 navíc používá termíny, které nejsou v korespondenci s termíny zavedenými a používanými v předchozí kapitole (reference point), a neuvádí zde ani žádný odkaz na upřesňující literaturu.
- problémy v textu:
 - a) *partByName* je několikrát použit, ale není nikde definován, navíc na str. 27 má 2 argumenty, na str. 29 a 57 už 3 argumenty.
 - b) str. 27, příklad 3 – jak je možné porovnávat *t_point* a *t_interval* pomocí \leq ?
 - c) str. 28, příklad 6 – uvedený časová specifikace je zjevně chybná - funkce *shift* nemá 3. argument.
 - d) str. 33, obr. 2.1 obsahuje buď několik chyb v anotačních výrazech nebo je nedostatečně vysvětlen, viz otázka k obhajobě c).
 - e) str. 38 – zřejmě pro větu účelovou je použit nesprávný anglický termín *aim clause*.
 - f) str. 63 – v textu se píše, že pro odvození pomocného uzlu se využívá i podfunktor. Ve vzorci (4.9) už ovšem podfunktor uveden není.
 - g) str. 73 podle textu se operátor $=\sim$ pojí s regulárními výrazy. Z příkladů na str. 73 a v sekci 5.4 je ale zřejmé, že se nejedná o regulární výrazy, ale o tzv. *divoké znaky* (*wildcards*), které mají slabší vyjadřovací sílu.

Otázky k obhajobě:

- a) str. 30 – jaký je přesně účel příznaku *distributive*? Vysvětlení v práci pokládám za nedostatečné.
- b) str. 31 – co by znamenalo umožnit spouštění nástrojů na PDT 2.0? Jednalo by se jen o překódování struktur 1:1 nebo by byly nutné další úpravy algoritmu? A co by bylo obecně potřeba pro použití nástrojů na zcela jiných vstupních datech (jiný formát morfologických značek, jiná struktura syntaktické anotace, případně jiný jazyk)?
- c) str. 33, obr. 2.1 co znamená $0.p$? Proč je $P.e = 2.s$? Pokud “postavit” předchází “prodat,” nemělo by místo $P.5.s \prec P.6.e$ být spíše $P.5.e \preceq P.6.s$?
- d) str. 61, sekce 4.3 – autor píše, že chybějící informace o podfunktorech je zásadní, proto podfunktory automaticky vygeneroval. Proč se tedy tato informace nemůže stejně automaticky doplnit až ve vlastním generativním algoritmu?

- e) str. 63 – při zpracování TG stromu se ignoruje pořadí poduzlů, na str. 10 se ovšem uvádí, že toto pořadí odpovídá tématu a rématu věty. Při generování tedy není zachováno aktuální členění věty?
- f) str. 75 – jak zvýší predikáty *testAttribute* a *testChildrenCount* vyjadřovací sílu dotazovacího formalizmu nad sílu predikátové logiky 1. řádu (podle str. 73)?
- g) str. 79 – jaká je lingvistická interpretace příkladu 3? Kde v popisu substitučního algoritmu je popsáno, že vzorce pod negací se při substituci ignorují?
- h) kap. 5 – v jakých konkrétních aplikacích se popsaný dotazovací a substituční algoritmus používá? Kolik uživatelů tyto aplikace mají? Jaké je v těchto ukazatelích srovnání s NetGraph pro PDT 2.0?
- i) sekce 6.1 – jsou uvedené rozdíly v anglických časech opravdu důležité z hlediska *generování*? Není skutečný rozdíl mezi generovanými variantami zanedbatelný co se týká “přijetí” věty čtenářem? Mimo jiné to naznačují i výsledky v sekci 6.2.2, které by jinak bylo nutné hodnotit jako špatné.

Závěrečné hodnocení:

Práce se zabývá výrazně aktuálními tématy z oblasti analýzy a generování textů v přirozeném jazyce. Výsledky práce vidím zejména v kvalitní implementaci několika programových systémů z dané oblasti, což vždycky vede ke zpřesnění a aktualizaci dosavadních teoretických algoritmů. Pro využití uvedených výsledků v praxi by ovšem ve většině případů bylo nutné výzkum ještě dále precizovat. Celkově konstatuji, že předložená dizertační práce **prokazuje** předpoklady autora k samostatné tvořivé práci a představuje vhodný podklad pro získání titulu Ph.D.

V Brně dne 1. listopadu 2007



Masarykova univerzita, Brno