Posudek diplomové práce

Matematicko-fyzikální fakulta, Univerzita Karlova

Autor práce: **Ján Faryad**
Název práce: **Extraction of multilingual valency frames from dependency treebanks**
Studijní program: Informatika
Studijní obor: Matematická lingvistika

Autor posudku: doc. RNDr. Markéta Lopatková, Ph.D. (oponentka)
Pracoviště: ÚFAL MFF UK

Téma práce

Předložená diplomová práce se věnuje problému získání valenční informace ze syntakticky anotovaných korpusů. Diplomant využívá závislostní korpusy anotované podle principů Universal Dependencies a navrhuje systém extraktorů, které z takto zpracovaných textů v daném jazyce extrahují informace o syntaktickém chování sloves tohoto jazyka a o jejich kombinatorickém potenciálu (autor je označuje jako valenční rámce). Základní verzi systému, která je jazykově nezávislá, mohou doplňovat extraktory pro konkrétní jazyky či jejich rodiny – autor vytvořil extraktory pro češtinu, slovenštinu a angličtinu. Dále diplomant navrhuje metodu, jak získané slovníky pro jednotlivé jazyky propojit a získat tak vícejazyčný slovník. Důležitou součástí práce je vyhodnocení jednotlivých extraktorů i výsledných slovníků a jejich porovnání s ručně vytvořenými evaluačními daty.

Obsah práce a její hodnocení

Práce se skládá – kromě krátkého úvodu a závěru – ze 3 obsáhlých kapitol, 5 textových příloh a bibliografie (celkem 123 stran), k dispozici je i příloha s daty, skripty, UDPipe modely a bloku "valency" pro Udapi.

První kapitola (Background, 32 stran) představuje základní koncepty, nástroje a data, která jsou dále využívána. Velmi oceňuji vhled do lingvistické problematiky, který autor získal a který ve své práci bohatě využívá, i množství uváděných příkladů. Lze zde najít drobné nedostatky jako některé příliš zjednodušující popisy či drobné nekonzistence, případně nepřesnou terminologii (aktanty bývají zaměňovány s argumenty, dichotomie aktant / adjunkt je též problematická). Za vážnější prohřešek považuji skutečnost, že autor v některých případech necituje původní zdroje (celý oddíl o FGP vychází ze shrnutí v disertační práci A. Vernerové, což je zde jediná citovaná práce).

Za vlastní jádro práce považuji ***druhou kapitolu*** (Monolingual valency frames extraction, 59 stran), ve které autor představuje architekturu celého (čistě pravidlového) systému extraktorů, podrobně pak popisuje tzv. extrakční jednotky (units) ošetřující jednotlivé jazykové jevy. Opět oceňuji důkladný a zasvěcený popis jednotlivých jevů a možností jejich zpracování. Pro vyhodnocení dopadu jednotlivých extrakčních jednotek autor vytvořil evaluační data (část PUD) a vlastní metriky zohledňující specifičnost úkolu, v textu pak úspěšnost řešení analyzuje. Nutno podotknout, že výsledné extrahované rámce lze těžko považovat za valenční rámce, neboť z povahy dat nelze

identifikovat jednotlivé významy sloves (a proto také příliš nedává smysl tyto extrahované rámce porovnávat s valenčními slovníky z rodiny vallex). K této kapitole mám několik dílčích dotazů, které uvádím níže.

Třetí kapitola (Cross-lingual valency frames linking, 16 stran) zkoumá tři metody propojení již extrahovaných slovníků pro dva jazyky. Tyto metody jsou založené na (i) slovním zarovnání paralelního korpusu, na (ii) podobné struktuře vět v UD a na (iii) podobnosti jazyků (na úrovni grafémů) – diplomant diskutuje vhodné rysy, které lze v těchto metodách použít, a hledání vah (parametrů) jak pro jednotlivé metody, tak pro jejich kombinaci. Pro evaluaci přitom opět využívá ručně vytvořených dat (k nim též otázku níž). Tato kapitola je méně vyargumentovaná, přínosy jednotlivých postupů jsou méně zřejmé, není jasný způsob volby parametrů pro testování (jde o náhodné generování a následnou evaluaci?), v řadě případů je popis poněkud zavádějící.

Přílohy A-C poskytují technické údaje týkající se využitých a/nebo vytvořených dat a nástrojů, struktury extrahovaných slovníků, formátů výstupu a instrukcí ke spuštění vytvořených nástrojů.

Jazyková úroveň textu a technické připomínky

Diplomová práce je psána srozumitelnou angličtinou, její úroveň však není příliš vysoká. Text navíc obsahuje větší množství jazykových chyb (zejména členy, nevhodná struktura věty apod.), v poslední kapitole se přidávají věty se špatnými vazbami či duplicitními slovesy, příp. též věty nedokončené – autor již zřejmě před odevzdáním nestihl jazykovou korekturu.

Drobné nedostatečnosti též vykazuje technické zpracování, zejména:

- špatné řazení bibliografických položek (de Marneffe patří pod M);
- chybějící seznam tabulek a obrázků;
- špatné/chybějící odkazy v textu (str. 18, 20, 21, 30, 32, 34, …);
- špatně vysázené tabulky (např. 2.7, 2.14).

Otázky k práci

- Z diskuse v odd. 2.1.1 mi není jasné, jak jste nakonec vyřešil problém se substantivy či adjektivy odvozenými od sloves (typy *zpívání*, *ukrytý*), které nejsou označeny jako *VERB* (upostag) – pokoušíte se pro ně extrahovat rámce? (Částečnou odpověď lze nalézt v odd. 2.7.5).
- K čemu se vztahuje # of frames/args/… v záhlaví tabulek 2.1, 2.3 atd.? Jde o celkový počet rámců/argumentů ve výsledném slovníku? Pokud ano, proč se u kategorie "obl" liší počty v tabulkách 2.9 a 2.11?
- **Mohl byste zde přiblížit ideu slučování rámců (odd. 2.4.6 Frame reduction)?**
- Oddíl 3.2.2 působí zmatečně – jaká data byla tedy použita pro evaluaci prolinkování? Mluví se zde o 100 větách z PUD, 100 větách z korpusu JRC Acquis a opět o 100 větách z PUD?
- Proč se práce v kapitole 3 omezuje na pravidlové metody? Co brání využití metod strojového učení?
- **Jak probíhá evaluace (cs-en) linkování vzhledem k vallexu?**

Závěr

Autor předloženou prací prokazuje, že se dobře orientuje v oblasti počítačové lingvistiky, zejména v otázkách souvisejících s jazykovými daty, extrakcí informací z korpusu a automatickým zpracováním textů. Vzhledem k jeho studijnímu programu lze vyzdvihnout i hloubku znalostí z lingvistiky a schopnost jejich využití při zpracování dat. Práce dále prokazuje, že diplomant je schopen analyzovat jazyková data a navrhnout a implementovat automatické nástroje pro jejich zpracování. Je si též vědom důležitosti podrobného vyhodnocování navrženého systému i jeho jednotlivých součástí.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze 31. 1. 2024

Markéta Lopatková