Diplomová práce - posudek vedoucího

Nguy Gang Linh: Návrh souboru pravidel pro analýzu anafor v českém jazyce

Předložená práce se zabývá automatickým určováním odkazování v česky psaném textu (analýzou anafor). Autorka měla za úkol se seznámit s problematikou práce, s dostupnými daty (Pražský závislostní korpus, PDT), a navrhnout a implementovat postup nebo několik postupů, jak odkazování v textu odhalit. Vyhodnocení úspěšnosti takového programu měla provést standardními metodami obvyklými v oboru počítačové lingvistiky.

Práce je rozdělena do Úvodu, čtyř hlavních kapitol a závěru; obsahuje rovněž čtyři přílohy a CD s programy a daty. V úvodu autorka popisuje daný problém (anaforu v češtině a její druhy). V další kapitole rozebírá možné postupy a podrobně popisuje předchozí práce. Ve 3. kapitole uvádí nezbytný popis dat PDT, která měla k dispozici a na jejichž základě měla dospět k řešení. Jádrem práce je kapitola čtvrtá (vlastní přínos autorky), kde popisuje vlastní řešení problému a jeho implementaci jednak pomocí algoritmu rozhodovacích stromů a nástroje C4.5, jednak pomocí autorkou vytvořených pravidel. V závěru shrnuje dosažené výsledky a uvádí přehlednou tabulku kvantitativně měřené úspěšnosti výsledků pomocí standardních měr (precision, recall). Rozsah práce je na diplomovou práci nadstandardní.

Hodnocení:

Autorka pracovala po celou dobu na diplomové práci samostatně, a její práce je jedním z prvních komplexních pohledů na zpracování a automatickou identifikaci anafory v češtině. Dosažené výsledky jsou kvantitativně v absolutních číslech velmi dobré až vynikající (byť je zatím není možno srovnávat s předchozími pracemi na češtině a PDT vzhledem k jejich absenci). Je ovšem podstatné, že z textu vyplývá, že autorka se zadanou problematikou zabývala velmi důkladně a z jazykového technického hlediska jí velmi dobře porozuměla (nejde tedy „jen" o slepé nasazení známého algoritmu na daný problém). Podstatným přínosem není tedy jen závěrečná úspěšnost a vytvořený automatický softwarový nástroj, ale i vlastní text, který jistě bude vhodným, až nutným čtením pro každého dalšího, kdo se touto problematikou bude dále zabývat. Je třeba ocenit i to, že přesto, že čeština není pro autorku rodným jazykem, to v práci není znát.

Práce je psána česky, je jasná a srozumitelná.. Formální požadavky práce splňuje, seznam literatury je dostatečný a relevantní (obsahuje práce vydané na pracovišti vedoucího i práce světové), obsah přiloženého CD je uveden.

Závěr: předloženou práci považuji po všech stránkách za práci splňující kritéria práce diplomové na MFF UK a doporučuji ji k obhajobě.

Praha, 28. 8. 2006

Jan Hajič, vedoucí DP, ÚFAL MFF UK