

Vyjádření vedoucího k diplomové práci

Josef Toman: Automatická anotace angličtiny na tektogramatické rovině

Josef Toman se ve své diplomové práci zabýval automatickou analýzou angličtiny, a to pro specifický účel zvýšení efektivity (snížení pracnosti a zvýšení přesnosti) manuální anotace anglických vět na tektogramatické rovině popisu jazyka, jak ji definuje Pražský závislostní korpus (verze 2.0).

Práce je členěna na devět kapitol a seznam použité literatury. V úvodní kapitole autor definuje cíle práce, účel a zvolenou metodiku. Ve druhé kapitole popisuje kontext, ve kterém práce vznikala – projekt Pražského anglického závislostního korpusu (PEDT). V další kapitole jsou popsány zdroje dat, ze kterých projekt PEDT vychází: Penn Treebank (vč. Nombanku) a nadstavba tohoto korpusu z BBN (zejména pojmenované entity a koreference). Ve čtvrté kapitole je popsána nedílná součást každé práce v experimentální části počítačové lingvistiky – evaluace a použité evaluační metriky. Pátá kapitola je vlastním jádrem práce (i když část autorovy práce, zejména párování tektogramatických stromů pro účely porovnání a evaluace, je uvedeno už v kapitole předchozí). V ní popisuje postupně všechny jazykové jevy (angličtiny), kterým ve své práci věnoval pozornost a psal pro ně příslušné moduly pro automatické zpracování. V šesté kapitole autor popisuje další část vlastního přínosu – kontroly v průběhu anotace (tedy „online“), a v kapitole sedmé pak další procedury, které se provádějí automaticky až po manuální anotaci (tedy už bez možnosti lidské korekce). Vyhodnocení vytvořených metod je pak popsáno v osmé kapitole a poslední kapitola pak tvoří závěr diplomové práce.

Hodnocení:

Práce je psána česky, je srozumitelná a autorův vlastní přínos je identifikovatelný (jedná se o práci a výsledky popsané v kapitolách 4-8). Objem vlastní diplomové práce je obvyklý pro diplomové práce na MFF. Srozumitelnost práce vhodně doplňují grafy a vizuálně vydařené příklady. Tabulky, zejména v evaluační části, jsou rovněž vhodně navrženy. Autor implementoval řadu modulů pro automatickou anotaci angličtiny, řadu z nich pak v českém prostředí i univerzálně zcela nových. Objem práce, který tyto moduly obsahují z programátorského hlediska a z hlediska evaluace, je nadprůměrný. Autor řádně vyhodnotil úspěšnost jednotlivých kroků i celé posloupnosti (kap. 5, kap. 8). Za jeden z nejzajímavějších výsledků v evaluační části lze považovat ověření vlivu automatického předzpracování na manuální anotaci a její přesnost.

Práci lze vytknout pouze některé formulace, které by v odborném textu neměly být, ať už jde o hovorová nebo příliš expresivní slova („žalostně“, „mizerný“). Dále autor vhodně užívá rozlišení na 1. osobu jednotného a množného čísla, jak píše v úvodu, ale některé obraty s použitím zejména 1. os. Jednotného čísla nejsou rovněž vhodné pro věcný styl. Jejich použití sice nečiní práci nesrozumitelnou (spíše naopak), ale současný standard je v zásadě nepřipouští.

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



Praha, 4. 5. 2009, Jan Hajič, ÚFAL MFF UK