Abstrakt

Název: Zpracování turkických jazyků
Autor: Sibel Ciddi
Katedra: Ústav formální a aplikované lingvistiky,
Matematicko-fyzikální fakulta, Univerzita Karlova v Praze
Vedoucí diplomové práce: RNDr. Daniel Zeman, Ph.D.
Abstrakt: Tato práce představuje a na příkladu turečtiny demonstruje několik metod morfologického zpracování vhodných pro turkické jazyky, jejichž počítačové zpracování přináší sadu specifických problémů. Přínosem práce je také značné rozšíření lexikální databáze a souvisejících dat potřebných pro morfologickou analýzu a syntézu; tato data jsou nyní volně dostupná veřejnosti. S ohledem na vysoce produktivní a aglutinační tureckou morfologii a s ní spojenou řídkost dat byl omezený rozsah slovníku významnou překážkou počítačového zpracování jazyka, zvláště pokud jde o zpracování statistickými metodami. Proto jsme důkladně otestovali a vyhodnotili veřejně dostupný, na konečných převodnících založený morfologický analyzátor TRmorph. Zaměřili jsme se na rozšíření záběru a slovníku tohoto analyzátoru. Za tím účelem jsme navrhli heuristické metody pro získávání pojmenovaných entit a víceslovných výrazů. Další vylepšení spočívá ve slovníkovém rozpoznávání víceslovných výrazů, které je předstupněm k jejich morfologickému zpracování. Výsledky experimentů ukazují, že takové heuristické rozšíření slovníku slibně zvyšuje pokrytí textů, které jsme s pomocí TRmorphu rozebrali. Slovníková metoda nejen umožňuje správnou tokenizaci víceslovných výrazů, ale také tím snižuje (lexikální) morfologickou nejednoznačnost a připravuje analyzovaný text pro použití v aplikacích vyšší úrovně. Námi navržený přístup tak zvyšuje účinnost čistě pravidlového morfologického analyzátoru s konečnými převodníky.
Klíčová slova: morfologická analýza, konečný převodník, konečný automat, rozpoznávání pojmenovaných entit, rozpoznávání víceslovných výrazů, morfologická a lexikální nejednoznačnost.