Pavel Češka: Vyhledávání v nesegmentované mluvené řeči

Posudek vedoucího diplomové práce

Diplomová práce Pavla Češky nese název „Vyhledávání v nesegmentované mluvené řeči" a zabývá se (přesněji) vyhledáváním tématicky relevantních pasáží v (tématicky) nesegmentovaných záznamech mluvené řeči. Je třeba zdůraznit, že v tomto oboru je to práce do značné míry pilotní a vychází z účasti autora na *Cross-Language Evaluation Forum* (CLEF) v roce 2007, konkrétně v sekci *Cross-Language Speech Retrieval Track*. Předběžné výsledky této práce byly publikovány ve sborníku CLEF a také v sérii LNCS vydavatelství Springer.

Předkládaná práce se skládá ze 7 kapitol na celkem 69 stranách a přílohy ve formě CD obsahující software vytvořený v rámci diplomové práce a ukázková data. První kapitola obsahuje úvod a motivaci pro celou práci, včetně specifikace jejích cílů. Ty spočívají především ve dvou bodech: jednak v přípravě nástrojů umožňujících generování parametrizovaných kolekcí z automatických přepisů mluvené řeči, jednak v provedení experimentů a nalezení nejvhodnějších metod a parametrů kolekce pro úspěšné vyhledávání. V druhé kapitole diplomant seznamuje čtenáře s různými technikami používanými pro vyhledávání informací v psaném textu (vyhledávacími modely, možnostmi předzpracování dat, rozšiřování dotazů apod.) a také evaluačními metodami (včetně testů signifikance). Ve třetí kapitole diplomant popisuje data, na kterých prováděl své exprimenty, tedy kolekci audio záznamů výpovědí svědků holokaustu (poskytnuté nadací VHF v rámci projektu Malach), které byly automaticky převedeny do psaného textu (na FAV ZČU v Plzni), soubor popisů témat obsažených v této kolekci, a také ručně identifikované pasáže (ve výpovědích) relevantní vůči jednotlivým tématům (získané na MFF UK v Praze). V kapitole čtvrté autor převádí problém vyhledávání v nesegmentované mluvené řeči na problém vyhledávání v jejích automaticky segmentovaných textových přepisech. V této kapitole je také popsána převzatá evaluační metrika. Jádrem práce jsou kapitoly 5 a 6 obsahující popis vytvořených programových nástrojů, tedy skriptů pro generování parametrizovaných kolekcí a evaluaci výsledků (kapitola 5), a popis série experimentů vedoucí k empirické optimalizaci použitých parametrů, zejména vhodnost či nevhodnost použití lematizace, stoplistu, ASR systému, délky a přesahu generovaných segmentů atd. (kapitola 6). V kapitole 7 autor stručně a přehledně shrnuje a diskutuje dosažené výsledky.

Práci je možné charakterizovat jako výzkumně-experimentální, přičemž množství exprimentů je poměrně velké. Pro jejich přípravu bylo nutné vytvořit software menšího rozsahu; nástroje pro lematizaci a pro samotné vyhledávání byly převzaty.

Hodnocení

Předkládáná práce plně splňuje požadavky kladené na diplomovou práci na MFF UK. Vyniká zejména v systematičnosti a korektnosti postupu při provádění experimentů (rozdělení testovacích a trénovacích dat, testování signifikance výsledků, jejich validace na testovacích datech apod.) a také v diskusi dosažených výsledků. Pokud bych měl práci něco vytknout, byla by to přílišná stručnost v některých pasážích mírně stěžující jejich pochopení (většinou však doplněná referencí na relevantní literaturu). Místy se objevují také překlepy a gramatické chyby, sporadicky i chyby faktické (viz např. matematické formule v sekci 2.3.2). Diplomovou práci doporučuji k obhajobě.

Vypracoval: Pavel Pecina, ÚFAL MFF UK

Praha, 3.9. 2008