Posudek diplomové práce
Pavla Češky

Vyhledávání v nesegmentované mluvené řeči

Karlova Univerzita, Matematicko-fyzikální fakulta

Obsah diplomové práce

Práce Pavla Češky se zaměřila na experimenty se systémy automatického vyhledávání informací modifikovaných pro vyhledávání v rozsáhlých zvukových datech, vyhledávány jsou relevantní pasáže pro zadaná témata. Autor musel řešit otázky související s nekvalitním automatickým přepisem řeči a specifickými vlastnostmi češtiny.

První a druhá kapitola jsou věnovány úvodu. Autor v nich popisuje cíle práce, nejrozšířenější algoritmy vyhledávání informací v textu, jejich existující implementace a postup vyhodnocení úspěšnosti jejich vyhledávání.

Třetí kapitola představuje v experimentech použité audio-vizuální nahrávky projektu MALACH. Kromě struktury korpusu, formátů zvukových dat a automatických textových přepisů, jsou popsány postupy anotátorů, kteří vyhledávali relevantní pasáže k vybraným dotazům zahrnutých do experimentů.

Modifikace problému vyhledávání v audiu na vyhledávání v textu je detailně popsána ve čtvrté kapitole. Pojem dokument je v tomto kontextu nadefinován jako časové okno v audio záznamu. Jsou zde popsány i nutné úpravy způsobu vyhodnocení takto upraveného vyhledávání.

Naprogramované nástroje pro přípravu automatických přepisů (kolekcí dat) i dotazů pro vyhledávací algoritmy a implementace jejich vyhodnocování jsou rozebrány v páté kapitole. Dobře komentované zdrojové kódy a ukázková data jsou k dispozici na přiloženém CD.

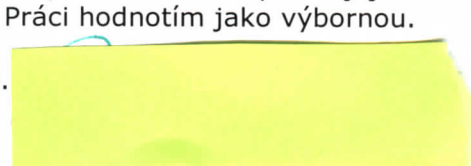
V experimentech se autor zaměřil na vliv lemmatizace, typu vyhledávacích algoritmů, různých délek a překryvů časového okna dokumentu, rozšiřování dotazu, výběru přepisu (kanálu), vylučování slov (stoplist) a zahrnutí detailního popisu dotazu do vyhledávání. Finální experimenty provedl na testovacích datech. Průběh experimentů, jejich statistické vyhodnocení a tabulky jsou v šesté kapitole.

Závěrečná sedmá kapitola přináší celkové hodnocení, ve kterém je vyzdvižen pozitivní vliv lemmatizace, normalizace hovorových slov a zvolení délky okna co nejblíže mediánu délky relevantních pasáží. Naopak autor ukázal, že překrývající se okna (dokumenty) snižují úspěšnost a vylučování funkčních slov úspěšnost neovlivňuje.

Hodnocení

Autor práce předvedl schopnost nastudování, presentování a aplikace nejnovějších jazykových informačních technologií a především schopnost jejich dalšího vývoje originálními a systematicky vedenými experimenty. Práci hodnotím jako výbornou.

Mgr. Nino Peterek, Ph.D.
UK MFF, ÚFAL



V Praze 28.8.2008