Posudek vedoucího diplomové práce

Michal Koktan: Automatické rozpoznávání (analýza) sentimentu

Diplomová práce se zabývá automatickým rozpoznáváním sentimentu[1]. Hlavní důvody zadání práce jsou dva: 1) použití výsledků práce pro výzkumné účely v oblasti automatického zpracování přirozené řeči v podobě textu na katedře KIV; 2) integrace implementovaných metod do prostředí České tiskové kanceláře (ČTK). Hlavní cíl práce je prostudovat existující metody automatického rozpoznávání sentimentu a stávající strukturu textových databází ČTK. Na základě této studie pak zvolit vhodnou metodu/y, navrhnout a implementovat systém pro automatické rozpoznávání sentimentu a ověřit jeho funkčnost.

Autor nejdříve popisuje problém analýzy sentimentu spolu s metodami rozpoznávání. Úloha rozpoznávání sentimentu je zde uvažována jako speciální případ úlohy automatické klasifikace dokumentů. Diplomant proto správně vybral tři metody, které jsou úspěšně používány v dané oblasti: Naivní Bayesův klasifikátor, Maximální entropie a Support Vector Machine. Následuje stručný popis metod paralelního zpracování textu, který obsahuje paralelní implementaci algoritmu Expectation Maximization. Zde by mohlo být metod více. Nicméně vzhledem k tomu, že se při klasifikaci paralelismus nepoužívá a dále bude ukázáno, že praktický přínos nemá, považuji tento popis za dostatečný.

V další kapitole je vysvětlen pojem korpus a stručně popsána problematika týkající se jeho tvorby v oblasti analýzy sentimentu. Následuje popis dat od ČTK. Dále autor uvádí přehled dostupných nástrojů pro klasifikaci. Kapitola také obsahuje stručný popis evaluačních metrik a použitou metodu křížové validace.

Dále bylo nutné vybrat český korpus vhodný pro analýzu sentimentu. Původně jsme předpokládali použití korpusu z Ústavu formální a aplikované lingvistiky Univerzity Karlovy, který nám ale bohužel nebyl poskytnut. Následné vyhledávání na Webu ukázalo, že v současné době není k dispozici žádný volně dostupný korpus anotovaný pro analýzu sentimentu. Diplomant proto musel vytvořit korpus sám. Pro tvorbu byla použita data dodaná od ČTK. Diplomant vytvořil anotační nástroj, pomocí kterého byl korpus označkován. Vytvořený korpus obsahuje hodnocení pěti osob, celkem se jedná o 750 odstavců označených dle sentimentu do tří tříd (pozitivní, negativní a neutrální). Nepředpokládaná tvorba korpusu je jedním z důvodů pozdního odevzdání diplomové práce.

V další kapitole se diplomant věnuje samotnému návrhu a implementaci systému pro automatickou klasifikaci sentimentu. Systém je založen na nástroji Minorthird, který nejlépe vyhovoval pro danou úlohu (open source, obsahuje všechny požadované klasifikační algoritmy, atd.). Nástroj Minorthird byl pro účely DP upraven. Hlavní úpravy spočívají ve změně formátu vstupních a výstupních souborů a v možnosti paralelního spouštění klasifikace. Následuje popis vykonaných experimentů a dosažené výsledky. Autor použil čtyři různé trénovací/testovací sady, které se lišily počtem zahrnutých dokumentů. Výsledky všech tří klasifikátorů byly srovnatelné, úspěšnost rozpoznávání se pohybovala kolem 70 %. Pro paralelní klasifikaci autor navrhl a implementoval dvě metody kombinace klasifikátorů: metodu lineární kombinace a většinového hlasování. Experimenty ukázaly, že použití paralelního zpracování pro danou úlohu je nevýznamné (zkrácení doby o 15 s a zvýšení přesnosti klasifikace asi o 1 %). Pro účely parametrizace je použita nejjednodušší metoda a to dokumentová frekvence. V práci bych uvítal popis dalších parametrizačních metod. Dalším přínosem by bylo zřejmě předzpracování textu.

Průvodní dokument je připraven v systému LaTeX. Má celkem logickou strukturu, jen v některých částech byla složitější orientace. Úvod se jeví jako neúplný (závěrečným časovým tlakem nebyla zřejmě použita finální verze). Práce obsahuje některé obtížně srozumitelné formulace, které vznikly nejspíše překladem anglické literatury. Dokument obsahuje několik nepřesností, což je způsobeno úrovní autorových znalostí v dané problematice. Hotové řešení během testování fungovalo bez chyb. Na přiloženém CD postrádám popisný readme soubor, který je ale uveden na konci práce v přílohách.

Diplomant byl aktivní. Nicméně bych při práci ocenil větší samostatnost. Některé úlohy, kde bych očekával zcela samostatné vyřešení, byly vyřešeny až po společné konzultaci.

Předložená diplomová práce přes uvedené nedostatky splňuje požadavky zadání. Autor prokázal, že dokáže inženýrským způsobem řešit zadané problémy. Práci doporučuji k obhajobě a hodnotím klasifikačním stupněm

„velmi dobře"

Ing. Pavel Král, Ph.D.
vedoucí DP

V Plzni 7. srpna 2012

Otázky: V práci uvádíte, že jsou nejlepší výsledky paralelní klasifikace (3 klasifikátory) 73 %, ale v tabulce 16 máte uvedenou přesnost 76 %. Vysvětlete prosím.

[1] Pojmem „sentiment" rozumíme názor (hodnocení) uživatele na daný objekt (případně na určitou osobu).