Diplomová práce – posudek vedoucího

Ondřej Kazík: Lingvistická komprese textu

V této diplomové práci autor navrhl kompresní metodu částečně využívající struktury komprimovaného českého textového dokumentu. Samotná komprese probíhá ve dvou fázích. V první fázi se zjistí pro každou větnou část v dokumentu její typ, tj. posloupnost slovních druhů, které se v ní vyskytují. Protože kompresní algoritmy by měli být rychlé, jedná se pouze o přibližné zjištění slovních druhů pomocí neuronových sítí, které se může od skutečnosti lišit, pro potřeby komprese je však dostačující.

V druhé fázi komprese se pak kóduje zjištěný typ větné části. Pro jednotlivá slova ve větě takto známe slovní druh a jejich zakódování tedy zabere méně místa, než kdybychom slovní druh neznali.

Kompresi vylepšuje velké množství staticky získaných inicializačních slovníku, jeden je pro typy větných částí, další jsou například pro jednotlivé slovní druhy. Tyto slovníky byly získány analýzou velkého množství dokumentů. V práci jsou představeny různé heuristiky pro vytváření těchto slovníků.

Domnívám se, že jde o velmi zdařilou práci, ve které autor prokázal, že umí nastudovat odbornou literaturu a na jejím základě navrhnout netriviální algoritmy. Samotný text práce má vysokou formální úroveň.

Experimenty pro srovnání algoritmů byly vhodně navrženy a získané výsledky jsou srozumitelně a přehledně vyhodnoceny. Na dokumentech střední velikosti autorův algoritmus dokonce překonal programy bzip2 a gzip. Domnívám se, že práci by bylo možno publikovat na Data Compression Konference.

Předloženou práci považuji po všech stránkách za práci splňující kritéria pro diplomové práce na MFF UK a doporučuji ji k obhajobě.

Praha, 25. 8. 2009

Mgr. Jan Lánský, Ph.D.