Posudek dizertační práce

Roman Chýla

Automated methods of textual content analysis and description of text structures

Motivací pro předloženou disertační práci byla snaha vytvořit systém, který by ve webovém prostředí implementoval „Universal Semantic Language" (USL) navržený a studovaný Doc. PhDr. Vladimírem Smetáčkem, CSc., vedoucím dizertační práce Romana Chýly v sedmdesátých letech 20. století. Práce je napsána v angličtině. Na základě uvedené motivace bylo stanoveno pět dílčích otázek, na které autor hledá odpověď:

1. Is the idea of the universal semantic language (USL) translatable to the form of a working application in a different domain?
2. Can we translate the texts from their natural form into the language of USL and actually avoid the problems of increasing entropy?
3. Can the process of extraction of facts work?
4. Is the current form of knowledge representation management sustainable? And what tools and methods one needs to develop to improve them?
5. Is it possible to have one, universal definition of meaning?

Práce obsahuje krom úvodu pět kapitol: II – Semantics and Knowledge Representation, III – Content Analysis, IV – Software for Content Analysis , V – Evaluation of SEMAN a VI – Conclusions. V kapitole II jsou charakterizovány různé prostředky pro reprezentaci sémantiky a znalostí a diskutován jejich vztah k USL. Velká pozornost je věnována otázkám významu, sémantickým polím a lexikální sémantice. Autor se zabývá i vztahem USL k výsledkům Goddarda a Wierzbicke. Na konci kapitoly II jsou diskutovány nevýhody a problémy USL. Rozsah kapitoly je 26 stran. Lze říci, že autor věnoval odpovídající úsilí studovaným otázkám. Domnívám se však také, že této kapitole, stejně jako celé práci, by prospělo zařazení přehledného popisu USL.

Kapitola III – Content Analysis začíná diskusí o různých aspektech analýzy obsahu. Poté jsou uvedeny čtyři typy obsahové analýzy – deskriptivní, inferenční, psychometrická a prediktivní. Je diskutován vztah projektu SEMAN k prediktivní analýze. Dále jsou přehledně uvedeny jednotlivé kroky procesu obsahové analýzy a podrobněji diskutovány některé jeho aspekty. Kapitola II má celkem 14 stran.

Jádrem práce jsou kapitoly IV a V. Kapitola IV – Software for Content Analysis se zabývá softwarem pro analýzu obsahu. V první části kapitoly IV je nejprve obecně pojednáno o softwarových systémech zabývajících se touto problematikou. Poté jsou vybrány tři z nich jako vhodní reprezentanti pro porovnání se systémem SEMAN, který je popsán v druhé části kapitoly IV. Jedná se o systémy General Inquirer, TABARI a Yoshikoder. Každému z těchto systémů je věnován jeden odstavec a každý z těchto systémů je podrobně popsán. Je stručně popsána historie každého systému, jeho důležité funkce a uvedeny ukázky výstupů Pro systémy TABARI a Yoshikoder je ještě v samostatném odstavci „Concluding remarks" uvedeno stručné zhodnocení systému. První část kapitoly IV obsahující popisy uvedených tří systémů má celkem 30 stran.

Druhá část kapitoly IV obsahuje popis systému SEMAN, který byl autorem vytvořen s cílem zodpovědět výše uvedené otázky. Pro popis systému jsou využity tři kroky prováděné při zpracování množiny dokumentů: (1) – předzpracování pomoci nástrojů NLP (natural langure processing), (2) – překlad do sémantických kódů, (3) – analýza a export výsledků. Jsou podrobně popsány použité metody a uvedeny i příklady výstupů. V posledním odstavci kapitoly IV je srovnání systému SEMAN s dalšími prostředky pro sémantickou analýzu. Popis systému SEMAN je celkem na 36 stranách.

Kapitola V – Evaluation of SEMAN obsahuje podrobné zhodnocení systému SEMAN. Rozsah kapitoly je 36 stran. První část hodnocení se týká porovnávání vzorů (pattern matching). Systém SEMAN je porovnáván se systémem BibClassify vyvinutým v CERNu. Chování obou systémů je velmi podrobně srovnáváno, je použit soubor 2084 náhodně vybraných textů z oblasti fyziky. Je testována schopnost systému SEMAN vyrovnat se se sémantickou nejednoznačností (semantic ambiguity). K tomuto účelu byla využita úloha automatické klasifikace dokumentů. Možnosti systému SEMAN byly srovnávány se systémem založeným na SVM (support vector machine). Ke srovnání jsou použity corpusy *20 Newsgroups* a *HEP* (High Energy Physics). Způsob porovnání je popsán velmi podrobně, jsou přehledně uvedeny výsledky porovnání a podrobně komentovány.

Kapitola VI – Conclusions shrnuje obsah práce a komentuje výsledky hodnocení systému SEMAN popsané v kapitole V. Přílohy práce obsahují mimo jiné i bibliografii původního systému SEMAN a webovou adresu, odkud je možné získat implementaci nového systému SEMAN.

Autor uvádí v úvodu, v odstavci 1.3, že práce obsahuje tři výzkumné výsledky (research contributions):

1. We have developed a new research tools.
2. We have reviewed the theory behind the USL and summarized its application into the context of content analysis.
3. We have evaluated the ability of the system to code texts and retrieve semantic relationships.

Na základě prostudování práce a svých znalostí problematiky souhlasím s tímto tvrzením, stejně jako s dalším tvrzením ohledně možností systému SEMAN uvedeným v odstavci 1.3.

K práci nemám podstatných připomínek kromě připomínky uvedené ke kapitole II. Domnívám se totiž, že práci by prospělo zařazení přehledného popisu USL v samostatné kapitole. To však v žádném případě nemění nic na tom, že tuto disertační práci považuji za plně způsobilou pro předložení k obhajobě. Doporučuji, aby na základě její úspěšné obhajoby byl Romanovi Chýlovi udělen titul Ph.D.

prof. RNDr. Jan Rauch, CSc.

V Praze, dne 28. 11. 2011

Posudek dizertační práce

Roman Chýla

Automated methods of textual content analysis and description of text structures

Motivací pro předloženou disertační práci byla snaha vytvořit systém, který by ve webovém prostředí implementoval „Universal Semantic Language" (USL) navržený a studovaný Doc. PhDr. Vladimírem Smetáčkem, CSc., vedoucím dizertační práce Romana Chýly v sedmdesátých letech 20. století. Práce je napsána v angličtině. Na základě uvedené motivace bylo stanoveno pět dílčích otázek, na které autor hledá odpověď:

1. Is the idea of the universal semantic language (USL) translatable to the form of a working application in a different domain?
2. Can we translate the texts from their natural form into the language of USL and actually avoid the problems of increasing entropy?
3. Can the process of extraction of facts work?
4. Is the current form of knowledge representation management sustainable? And what tools and methods one needs to develop to improve them?
5. Is it possible to have one, universal definition of meaning?

Práce obsahuje krom úvodu pět kapitol: II – Semantics and Knowledge Representation, III – Content Analysis, IV – Software for Content Analysis , V – Evaluation of SEMAN a VI – Conclusions. V kapitole II jsou charakterizovány různé prostředky pro reprezentaci sémantiky a znalostí a diskutován jejich vztah k USL. Velká pozornost je věnována otázkám významu, sémantickým polím a lexikální sémantice. Autor se zabývá i vztahem USL k výsledkům Goddarda a Wierzbicke. Na konci kapitoly II jsou diskutovány nevýhody a problémy USL. Rozsah kapitoly je 26 stran. Lze říci, že autor věnoval odpovídající úsilí studovaným otázkám. Domnívám se však také, že této kapitole, stejně jako celé práci, by prospělo zařazení přehledného popisu USL.

Kapitola III – Content Analysis začíná diskusí o různých aspektech analýzy obsahu. Poté jsou uvedeny čtyři typy obsahové analýzy – deskriptivní, inferenční, psychometrická a prediktivní. Je diskutován vztah projektu SEMAN k prediktivní analýze. Dále jsou přehledně uvedeny jednotlivé kroky procesu obsahové analýzy a podrobněji diskutovány některé jeho aspekty. Kapitola II má celkem 14 stran.

Na základě prostudování práce a svých znalostí problematiky souhlasím s tímto tvrzením, stejně jako s dalším tvrzením ohledně možností systému SEMAN uvedeným v odstavci 1.3.

K práci nemám podstatných připomínek kromě připomínky uvedené ke kapitole II. Domnívám se totiž, že práci by prospělo zařazení přehledného popisu USL v samostatné kapitole. To však v žádném případě nemění nic na tom, že tuto disertační práci považuji za plně způsobilou pro předložení k obhajobě. Doporučuji, aby na základě její úspěšné obhajoby byl Romanovi Chýlovi udělen titul Ph.D.

prof. RNDr. Jan Rauch, CSc.

V Praze, dne 28. 11. 2011