Posudek vedoucího diplomové práce

Jméno a příjmení autora posudku:	RNDr. Pavel Pecina, Ph.D.
Jméno a příjmení autora práce:	Sara F. van de Moosdijk
Název práce	Minig texts at the discourse level

Text posudku

The main topic of the thesis is knowledge extraction from textual documents. The applied approach, based on analysis of linguistic discourse, is quite novel and challenging. It goes beyond the traditional methods, which usually treat documents as bags of words, and attempts to model relationships between larger units of texts to obtain more reliable information. The method exploits linguistic discourse and Formal Concept Analysis (FCA) and is applied to documents from the medical domain.

The thesis is written in English, the main text spans 53 pages and is split into 6 chapters plus bibliography and attachments on additional pages. After a brief introduction in Chapter 1, the author describes the theoretical background of the method in Chapter 2 and the data collection in Chapter 3. The main contribution of the work is described in Chapter 4 focused on extracting discourse relations and Chapter 5 devoted to representation using FCA and knowledge detection. The work is concluded in Chapter 6.

The thesis is well-structured. The theoretical background (of both linguistic discourse and knowledge discovery) is well explained and presented to a reader. The main contribution of the work includes 1) a modification of a method for extraction of relations from medical text (the original method was applied on the test corpus and based on manual evaluation the method was improved to provide better results) and 2) a method for converting the discourse relation data to FCA structures (including definition of pattern structures and exploitation of external resources – UMLS) and 3) an algorithm to process the structures and generate the final result (closed pattern concepts). The entire pipeline was applied to a corpus of 600 articles from the medical domain (obtained from the PubMed database).

My only comment is related to the evaluation of the results. The first step of the procedure was empirically evaluated (and the method was improved based on the results), but the evaluation of the next steps (Section 5) is not that thorough. I fully realize that such evaluation would difficult (it would require a lot of manual effort and also deep knowledge of the domain), but it would be nice to see some performance estimation of the method or its deeper comparison with the traditional approaches.

Sara F. van de Moosdijk developed a complex system for extraction of knowledge from domain-specific texts based on analysis of linguistic discourse and application of Formal Concept Analysis. The presented work certainly is a good starting point for further research and I recommend the thesis to be defended.

Doporučení k obhajobě

Z výše uvedených důvodů práci *doporučuji* k obhajobě.

Soutěž studentských prací

Vynikající práce vhodná soutěže studentských prací: **NE**.

V Praze dne 1. 9. 2014

Podpis: