Posudek oponenta diplomové práce

Jméno a příjmení autora posudku: Vladislav Kuboň

Jméno a příjmení autora práce: Ilana Rampula

Název práce: Semantic relation extraction from unstructured data in the business domain

---

Vlastní text (sem prosím napište text posudku, délka textu posudku není omezena):

The thesis tries to answer a question what kind of semantic information can be extracted from a large collection of unstructured business data and which methods are suitable for this endeavor. It consists from four main chapters, an introduction and a conclusions section. No data carrier is attached to the thesis.

Actually, although there is a special introductory section at the beginning of the thesis, it may be claimed that about a half of the thesis text constitutes a very thorough introduction into the problem. The whole chapter 1 deals with all aspects of the problem of extracting information from unstructured business data. The author introduces (at a great length) the task, the algorithms and methods used for the task and the evaluation measures. The main problem of not only this part of the thesis is the author's inability to segment the text into paragraphs of a reasonable length. It is sometimes even the case that one paragraph is longer than a full page (as, e.g., the first paragraph of thesis at page 1, a paragraph on pages 18 and 19 etc.). This lack of structure makes reading the thesis very difficult. Also the choice of words used by the author in some places is a bit non-standard (as, e.g., using the correct but very rare formulations such as "Craven and Kumlien (1999) coined such data instances weakly labeled.").

The second chapter describes the data and operations with them. It mostly only introduces existing tools, but later it also brings some new and original information. It describes and analyzes the nature of the business data available and reveals the surprising fact that although there were almost 2 millions of data items, they actually contain a negligible amount of semantic information which may be extracted. This is actually one of the very important results of the thesis – a clear proof of the fact that although unstructured data are nowadays very popular, the actual amount of information which they provide may be extremely limited. The author has discovered that the only semantic relation that may be extracted from the texts, is the relation Country-of-origin (inconsistently described as Country-of-birth in Fig.6), and that only about 73 000 text items actually contain a different country than Czech Republic.

The third chapter describes the methods used in experiments. The description contains everything important, it is well structured and clear. It clearly distinguishes between the original work performed by the author and a simple application of existing tools and methods. The last chapter presents all results obtained in the experiments. They are discussed in detail, the author shows that she knows very well what she was doing, she can correctly interpret the results and explain them. The results show that one of the two methods, the Distant supervision one, is clearly better than the other one. The author thoroughly discusses the reasons for this result, especially the drawbacks of the Snowball method. This part of the thesis is very well done.

The thesis ends with conclusions, very rich bibliography and lists of tables, figures and abbreviations. It also has one attachment of more or less technical nature.

The overall impression from the thesis is positive. The author was able to apply wide variety of existing tools; she obtained and analyzed a very unique set of business data, identified important features of the dataset, performed several experiments with various settings and correctly and thoroughly analyzed the results. The negative aspects concern solely the text of the thesis. Some of them have already been mentioned above, among the others are probably most important missing or late explanations of important notions such as positive and negative examples – they are for the first time mentioned at page 31 without any explanation. It comes as late as on page 41. Some choices made in the process of experiment preparations are also not explained (page 20 "Actually, this is one of the reasons we chose to use VW." – what are the other reasons?).

The technical quality of the thesis is also good, with some minor flaws. The text contains very few spelling or grammatical errors (as, e.g., ".. the score with be…" at p.26 or "… if we the recall is …" at p.54). Part of the text on page 22 is grey, the references to figures from the text are not consistent starting from Fig.6 (it seems that an original Fig.6 has been removed and the references were not updated afterwards). Also the reference to the content of the Table 19 at p.54 should have mentioned experiment 17 instead of 16.

All in all, the thesis represents a reasonable amount of solid scientific work and I recommend it for the defense.

Doporučení k obhajobě:

Z výše uvedených důvodů práci *doporučuji* k obhajobě.

Vynikající práce vhodná pro soutěž studentských prací	ANO ☐

Seznam soutěží studentských prací, viz http://www.mff.cuni.cz/studium/bcmgr/prace/

Pokud jste výše zaškrtli ANO, zdůvodněte prosím svůj návrh, případně uveďte konkrétní soutěž, pro kterou je práce vhodná (rámeček lze nechat prázdný, pokud za dostatečné zdůvodnění považujete text posudku):

V Praze dne: 1.6.2016

Podpis:**