Posudek *oponenta* diplomové práce

Jméno a příjmení autora posudku: Michal Novák

Jméno a příjmení autora práce: Sara van de Moosdijk

Název práce: Mining texts at the discourse level

Vlastní text:

Thesis description

The aim of the reviewed thesis is to bridge the fields of Knowledge Discovery in Databases and discourse-oriented NLP, using pattern structures (an extension of Formal Concept Analysis) to mine discourse structures in texts. The author applies this method to texts from the medical domain.

The thesis is structured as follows: An introduction in Chapter 1 shows the possible merit of taking discourse into account for text mining. Chapter 2 focuses on the theoretical background related to the two fields author attempts to bridge. Building a corpus to be used for experiments is the topic of Chapter 3. Chapters 4 describes a method for extracting discourse relations, while in Chapter 5, the author uses the extracted relations to build a representation of text using pattern structures. Conclusion, Bibliography and Appendix wrap up the thesis. The thesis consists of 61 pages including Bibliography and Appendix. No CD is attached.

Comments

As regards the form, the thesis is written in very good English, clearly and comprehensibly. One of the few shortcomings regarding the language is that the punctuation is sometimes missing, especially for commas that should follow “However” at the beginning of the sentence.

The terminology used in the work is well explained and used consistently.

Tables on page 31 and 36 showing the frequency of different error types for various relation types would be easier to comprehend if a row with totals per error type was added.

All in all, I greatly appreciate the style, which makes the thesis easy to read and understand, particularly by the extensive use of clarifying examples.

As for the content, Chapter 2 makes a very clear and sufficiently deep insight into the topic of discourse representation and the method of Formal Concept Analysis. However, the space devoted to the former seems to be too large considering the fact that in the end, the author abandoned the described formal discourse models (rightly, in my opinion) and decided to employ much simpler pattern matching. Instead, this space should have been used to investigate more the related works on text mining or discourse modeling in the medical domain. I have found multiple papers that closely relate to the presented topic (e.g. Prasad et al.: The biomedical discourse relation bank (2011); Mihăilă et al.: BioCause: Annotating and analysing causality in the biomedical domain (2013); Hahn et al.: Mining the pharmacogenomics literature—a survey of the state of the art (2012)).

As far as I am concerned, Chapter 4 is the strongest out of the parts describing the author's own work. Although the pursued approach of matching discourse patterns is not novel, the reader is presented with a thorough error analysis of the proposed method, supported by clarifying examples. Furthermore, the analysis serves as a basis for a set of refinements that eventually result in a quality improvement in discourse relations identification.

However, I have a remark regarding discourse relation extraction. Even though this practice becomes quite common among authors, I would avoid calling the chosen approach “unsupervised”. Unlike the works the author took her inspiration from, there is no machine learning technique applied after the extraction of the relations using patterns (or rules, heuristics).

I find Chapter 5 to be the most problematic one. I do not agree with the author's claim that the lattice of concepts representation is easy to interpret. Due to the overgenerating nature of the MetaMap tool, it is almost impossible that the convex hull of term sets of any two relations would not contain the top-most element in the ontology and thus, all the nodes on paths from the given terms to the top-most element.

In addition, the current implementation mixes the left-hand sides of the “[BOS …][because … EOS]” and “[BOS Because … ,][… EOS]” patterns together. While the left-hand side contains the consequence in the former pattern, this position is occupied by the cause in the latter one. The same holds for the right-hand sides as well as for the other pattern pairs.

The author faced performance issues while building the lattice representation and suggested running the algorithm in parallel or decreasing the number of terms associated with a text segment to rectify this problem. However, I doubt that this would help if we take an increased number of relations (or documents) into consideration. The reported time complexity of the CloseByOne algorithm is $O(|G|^2|D||L|)$, where G is a set of objects (relations), D is a set of patterns (terms) and L is a set of concepts in a lattice. On page 44, the author says that using the tree-like ontology to compare concepts will always result in lattice being a power set of G. If I understand it correctly, the complexity of CloseByOne algorithm is $O(|G|^2|D||2^G|)$ in this case, i.e., exponential with respect to the number of relations (or documents). Unless an effective pruning method is introduced, this makes the algorithm practically unusable.

Conclusion

Even though I doubt that the suggested approach is applicable to larger sets of relations or documents, the work satisfies the requirements set on a master thesis. Thus, I recommend that this thesis is accepted 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/prvzoryace/

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: 2.9.2014

Podpis: