Giang Binh Tran: Combining of text-based semantics and vision-based semantics

Review

The reviewed master thesis presents a novel approach to creating semantic models using a combination of traditional text-based features and visual features extracted from labeled images. The goal of this thesis is to introduce visual features as a useful addition to semantic modeling, develop an algorithm to combine them with traditional text-based features, implement a working semantic model based on both text- and image-based features, and finally, evaluate this model on standard computational linguistic tasks.

The reviewed work is structured as follows: The author explains aim and objective of the thesis (chapter 1), continues with related literature review (chapter 2) and explains basics of image processing (chapter 3). The core of the thesis lies in multimodal semantic model introduction (chapter 4) and experiments and evaluation description (chapter 5). Presented work contains an attachment with implementation.

The main contribution of this work is an original idea to extend standard text-based semantic models with visual features and an extensive evaluation of the proposed model on multiple semantic tasks. In evaluation on multiple traditional datasets (WordSim353, RG, etc.) and in multiple evaluation measures, the proposed multimodal semantic model performs very well in comparison with state-of-the-art systems.

One of the most questionable points of the evaluation, however, is a certain confusion about the coverage of the results. It is not clear from the text, whether the results are reported on the whole datasets (WordSim353, RG, etc.) or only on the selected subsets (e.g. on p. 40: "In our version of RG test, we have 47 pairs covered by the models, covering about 73% of the full RG test."). If that is so, a comparison against state-of-the-art methods is unsound. The Table 5.5 suggests a comparison of the same systems, while there are different datasets underlying. A valid evaluation is to report the results on the whole datasets because in computational linguistics, any method has to deal with uncovered (unknown) testing data.

Furthermore, in Table 5.5 comparing the WordSim353 Spearman coefficient, the author does not mention the best known score on this dataset published in [Agirre 2009], which is 0.78 (see Table 9 of [Agirre 2009]) and not 0.66 as claimed in the thesis. Also, Table 5.6 refers Pearson coefficient for RG dataset, which prevents comparison to state-of-the-art systems published in Spearman coefficient ([Agirre 2009]).

Another drawback of this work is unfortunately more than neglectable amount of grammatical mistakes and author's rather liberal approach to citations. About 50% of the references cited in the text are not listed in the bibliography section. In the bibliography section itself, the citations are not unified, the authors being cited in full names or shortcuts and in mixed bibliography styles (dots, commas and semicolons). Otherwise, the language used is easy to read and the grammatical mistakes do not prevent the reader from understanding the text.

To conclude, the presented work fulfills requirements to master thesis in both extent, scientific contribution, complexity and implementation. I highly recommend this work for accepting as master thesis.

References

[Agirre 2009] Agirre, E., Soroa, A., Alfonseca, E., Hall, K., Kravalova, J., and Pasca, M. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In Proceedings of annual meeting of the North American Chapter of the Association of Computational Linguistics (NAACL), Boulder, USA.

Prague, August 15, 2011

[redacted]