Pavel Pecina: Lexical Association Measures (Collocation Extraction)

This thesis builds significantly on a large body of work on collocation extraction, in implementing a wide range of association measures from variol sources, evaluating them over a total of 4 datasets (with different languages, corpus sizes and representations), and then combining the methods together in novel ways using supervised learning techniques. Notable aspects of this thesis are its tremendous breadth and depth, and empirical thoroughness. The main contributions of the research are: the clear documentation of a robust framework for collocation extraction (not novel in itself, but the attention to detail and clear argumentation makes it a gold-standard recipe for collocation extraction evaluation); development of a number of new datasets for collocation extraction experimentation; confirmation of the fact that different association measures perform differently over different datasets; and, most significantly, that it is possible to greatly improve collocation extraction performance by combining together multiple methods using supervised learning. This final contribution opens the way for a new approach to collocation-extraction meta-learning, and exploration of the interaction between individual methods.

The amount of material that the thesis pulls together is tremendous, leasing no stone unturned in terms of the range of collocation extraction methods it experiments with, and no room for misinterpretation of the empirical results it presents. Empirically, I found it flawless, while maintaining a feeling of effortlessness in the various pieces of empirical machinery that was brought to bear. I have provided Pavel with a large number of suggestions for low-level editorial changes to the thesis, to polish the presentation. I am konfident that he has made these changes in the submitted version of the thesis, and that the thesis is even more readable as a result.

If I were to have to ask for expansion over what is already in the thesis it would be in three areas. First, I would have liked to have seen slightly more analysis of the computational complexity efficiency of the different methods (mentioned in passing in Chapter 5), in terms of weighing up the "cost" of algorithms with their relative utility. Second, I would have liked to have seen side experiments looking at combining different preprocessing strategies for a given corpus (e.g. PDT), rather than looking at individual subtasks based on simple word + POS sequential information or alternatively dependency tuples. Finally, I would have liked to have seen more reflection on the complementarity between individual collocation extraction methods, in terms of the results of the meta-learning vs. the theoretical foundations of the individual measures. In this same vein, the thesis perhaps lacked punch slightly in its conclusion, in that I found myself wanting to see a definitive conclusion on what meta-learning method performs most consistently in the face of vast fluctuations in the performance of individual collocation extraction methods over different datasets. The final conclusion had a tinge of "different meta-learning methods work best in different contexts", mirroring the unpredictability of the underlying collocation extraction methods. In practice, the variance between the individual meta-learner methods was much less pronounced, however, and I believe that it should be possible to establish a preferred learning "formula" which works best on average, and where it can be shown empirically that even when it is not the best-performing method for a given dataset, it is within a relatively small margin of the best-performing method. In this way, the proposed strategy would have come across as less "ad hoc" in terms of what works well under what circumstances.

These suggestions, however, are more pointers to possible future work than requirements for extra work for the thesis to be worthy of acceptance in its current form. As stated above, the thesis is tremendous in its coverage, thoroughness and significance as is, for which Pavel should be heartily congratulated.

Timothy Baldwin Ph.D.

University of Melbourne, Carlton, Australia