UNIVERSITY OF
CAMBRIDGE

Anna Korhonen
Professor of Computational Linguistics
Co-director of the Language Technology Laboratory

November 4, 2017

Preliminary thesis report for

Stephen McGregor – Geometric Methods for Context Sensitive Distributional Semantics

This thesis introduces novel methodology for distributional semantics. Building on the idea that lexical semantics is context sensitive, the methodology aims to generate ad hoc semantic relationships in response to a linguistic or conceptual situation. Preserving the connection between the individual dimensions of word vectors and statistics pertaining to observations in text corpora, it is specifically designed to enable empirical exploration of distinctions between various semantic phenomena. It is largely this property that sets it apart from mainstream vector space models of distributional semantics - those that typically involve the factorisation of co-occurrence matrices or the incremental learning of representations using neural networks.

This is clearly a strong thesis. It introduces interesting, original ideas that make a real contribution to distributional semantics. The novel methodology is generally well-justified and technically solid. The experimental evaluation shows a clear benefit for semantic tasks in NLP.

I have a number topics that I would like to discuss during the viva, but I don't expect these (or the discussion) to lead into anything else but minor or straightforward modifications of the thesis. Here is the summary of the main points (I will leave the smaller and more specific issues for the viva):

- The research presented in the thesis has been published in multiple papers. Interestingly, although the thesis argues that the new methodology further develops and complements computational linguistic work on distributional semantics and although the task-based evaluations focus on NLP, none of the papers have been published in main NLP conferences. This is not a problem, but it raises an interesting question: how has the computational linguistics community reacted to this work? The specific question that I have in mind is: does this work solve a problem that has been recognised by the community as important for the further development of the area?

- A lot of effort has been put into (and space devoted to) explaining the theoretical background of the methodology. This is clearly very challenging because the methodology has been inspired by theories in multiple fields (philosophy, linguistics, cognitive science, information science, computational linguistics, among others). I find the first chapters would be much easier to read if the work was clearly related to a specific field (or just two, if needed) and if the other relevant fields were brought in as adjacent or related fields. If this work primarily develops and builds on the long line of computational semantics research within computational linguistics, I'd like to see more discussion on that line of research and less on others. For example, the vector space models are just one way of presenting semantics and the neural approaches are all very recent. To fully understand where we are now and where we ought to go next, it is important to look into what we had prior to these models and what alternative models might have offered.

- Related to the above point, the terminology used in the thesis (e.g. context, situation, concept) has different meanings in different fields and it is not fully clear which meanings are being assumed until we're on pages 30-40 or so (e.g. I found "context" was explained properly on p. 52 only). The glossary in the beginning of the thesis is great but is very brief and doesn't solve this problem.

- The one aspect of the methodology which is not explained very well is data. The generation of meaning is supposed to be dynamic and contextual - yet it operates on a corpus which is static / fixed?

- There are some other aspects of the methodology and evaluation that are not very well explained or justified in the current write-up (I will leave the examples for the viva)

- The methodology is evaluated in the context of several semantic NLP tasks, including word similarity and relatedness, metaphor and metonymy detection, and analogy completion. The results are mostly convincing. While the task-based evaluation is adequate for the thesis, these are not tasks with an end user. What about the usefulness of the approach for real-life application tasks known to benefit from semantics? Examples would be search, question-answering, text mining and dialogue, among others. There is little discussion on this in the thesis. The potential future applications mentioned (digital humanities and cognitive modelling) are valuable but they are outside the scope of mainstream NLP. They also suffer from small data which is known to limit the usefulness of statistical NLP, so how well would the method perform on them?

Anna Korhonen

Anna Korhonen