UNIVERSITY OF CAMBRIDGE

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

November 10, 2017

List of minor corrections for the thesis

Stephen McGregor – Geometric Methods for Context Sensitive Distributional Semantics

General corrections:

Explain, for all the experiments, whether you calculated statistical significance and whether the improvements or differences reported are significant.

Explain, for all the experiments to which this applies, how you did qualitative analysis: did you analyse all of the output data or only a subset? Did you do the analysis in some systematic manner?

Explain how you came up with the thresholds, the number of dimensions (20, 50, 200 etc) and sizes of windows used in your experiments. Did you just invent them or are they based on some preliminary experiment and then adopted for the rest of the thesis?

This is an optional correction, but I would recommend highlighting the best results in the tables for easier readability.

Specific corrections:

p. iv
Glossary: Because the terminology is quite cross-disciplinary in this thesis, I would expand this glossary a bit, give a more comprehensive definition for each concept and even mention, for concepts that are used in a non-conventional (from NLP perspective) sense, which field the definition comes from. I would also add “meaning” and “situation(al)” in this glossary.

Chapter 1

- Explain more clearly that this thesis belongs to / advances primarily the field of computational linguistics.

- Define more clearly what the thesis does and doesn't do: it develops computational linguistic methodology and evaluates it in the context of NLP tasks. Although based on theoretical insights from other fields (and although potentially useful for several fields), cross-disciplinary investigations are not included but are left for future work.
- P. 10 typo: there are two "the" words in the 3rd line of the 3rd paragraph

Chapter 2

- Section 2.2. When you start talking about concepts here, please define the intended properly and/or refer to the places in the thesis where they are defined properly (and mention the existence of that Glossary)
- Section 2.3 I found all this background on metaphor was a little out of place in this section. Consider moving the background elsewhere if possible? p. 24 This is the place where I would mention that no attempt is made in this thesis to discuss the conceptual structure (of the human brain).
- Section 2.4. Show some awareness of the long history of semantics research in NLP by including a paragraph or two that mention the main lines of research prior to / alongside VSMs. Then say explicitly that you will focus your literature review literature on the distributional semantics and neural approaches only. On p. 28 define "generalizability" better (it's a term with many meanings).

Chapter 3

- Section 3.2. p. 37 The concept "situation" should be defined better. In particular, what does a "situation of words in a large corpus" mean? p. 38 "Context" is defined here, too late in the write-up.

Chapter 4

- Section 4.1 The second paragraph talks about the cleaning process. Mention who performed this process.
- Section 4.2 There's a typo in the caption of table 4-7: delete one "the"

Chapter 5

- Here or earlier: Explain how you chose the specific tasks in chapters 5, 6, and 7 for your evaluation?
- p. 98 Here or elsewhere in the chapter (or in the future work section at the end of the thesis) discuss how factors such as part-of-speech (nouns, verbs, adjectives), polysemy and abstract vs. concrete, among others, many also influence your results alongside the obvious issue of frequency.
- P. 130 The end of the first paragraph on this page is difficult to understand – explain what you mean by "too conventional"

Chapter 6

- p. 139 BNC is a balanced corpus so shouldn't be colloquial in nature. Explain, if you can, how the 2000 sentences were selected.

Chapter 8

- Section 8.2. If possible, I would try and discuss the potential usefulness of the methodology introduced in this thesis for NLP at large, and for real-life applications (search, QA, etc). I would also discuss what it would take to make your methods useful for research in cognitive science.