Full set received on date 20/11/17

RECEIVED
15 NOV 2017
Partially rcvd.

Queen Mary
University of London

Outcome of Research Degree Examination form (RD03)

Name of candidate: STEPHEN McGREGOR
(*please underline surname*)

Degree for which examined: PhD

School/Institute: Elec. Engg. + Computer Sci.

Name of examiner one: JOHN BARNDEN

Name of examiner two: ANNA KORHONEN

1. We report that we have examined the thesis submitted by the candidate and have also examined him or her orally on the subject of the thesis and on subjects relevant to it.

Date of oral examination: 6 NOV. 2017

2. We report that we have determined one of the following outcomes:
(*complete only the relevant section, tick one box*)

Examinations for the PhD degree

a) Pass: the candidate has satisfied us in the examination for the degree ☐

b) Pass subject to minor amendments: to satisfy us in the examination the candidate must complete specified minor amendments within three months ☑

c) Pass subject to major amendments: to satisfy us in the examination the candidate must complete specified major amendments within six **or** nine months ☐

Please specify number of months here: ☐

d) Not pass: the candidate is permitted to re-enter for the examination and to re-present the thesis in revised form within eighteen months. ☐
(Is there to be a second oral examination? | Yes / No / Undecided | *please circle one)*

e) Not pass: the candidate is allowed to submit to a further oral examination within eighteen months ☐

f) The candidate has met the requirements for the degree of MPhil and should be awarded this degree ☐

g) To satisfy the requirements for the degree of MPhil, the candidate is required to make specified minor amendments to our satisfaction within three months ☐

h) The candidate is permitted to enter the examination for the degree of MPhil ☐

and to re-present the thesis in a revised form within twelve months

i) Fail: the candidate may not re-enter for the degree of PhD or MPhil ☐

Examinations for the MPhil degree

a) Pass: the candidate has satisfied us in the examination for the degree ☐

b) Pass subject to minor amendments: to satisfy us in the examination the candidate must complete specified minor amendments within three months ☐

c) Pass subject to major amendments : to satisfy us in the examination the candidate must complete specified major amendments within six or nine months or within a period specified by the examiners ☐ *Please specify number of months here:* ☐

d) Not pass: the candidate is permitted to re-enter for the examination and to re-present the thesis in revised form within twelve months ☐
(Is there to be a second oral examination? | Yes / No / Undecided | *please circle one)*

e) Not pass: the candidate is permitted to take a written paper or practical examination ☐

f) Not pass: the candidate is allowed to submit to a further oral examination within twelve months ☐

g) Fail: the candidate may not re-enter for the degree of MPhil ☐

Examinations for the MD(Res) degree

a) Pass: the candidate has satisfied us in the examination for the degree ☐

b) Pass subject to minor amendments: to satisfy us in the examination the candidate must complete specified minor amendments within three months ☐

c) Pass subject to major amendments : to satisfy us in the examination the candidate must complete specified major amendments within six or nine months or within a period specified by the examiners. ☐ *Please specify number of months here:* ☐

d) Not pass: the candidate is permitted to re-enter for the examination and to re-present the thesis in revised form within eighteen months ☐
(Is there to be a second oral examination? | Yes / No / Undecided | *please circle one)*

e) Not pass: the candidate is allowed to submit to a further oral examination within eighteen months ☐

f) Fail: the candidate may not re-enter for the degree of MD(Res) ☐

Will send to Research Degrees Office.
rcvd 15/4/17

3. Preliminary reports

Please attach your independent preliminary reports, signed and dated, to this form. Except in the case of request for a review of the decision of examiners, the candidate will not receive copies of the preliminary reports unless you indicate by ticking the box below that you want him or her to do so.

We wish the candidate to receive copies of our preliminary reports (*please tick)* ☑

4. Final joint report

The final joint report should **give the grounds on which your decision is based**. It should include the candidate's name, thesis title, the signatures of each of the examiners and the date.

The candidate will be provided with a copy of the final report. Please do not make reference to the preliminary reports in it unless you have indicated above that the candidate should received copies of them.

5. Amendments

If you have determined that you require the candidate to make specified amendments, please select one of the following options for completion and checking of amendments:

a) Where (minor) amendments are typographical only they may be made by hand on the originally-submitted copies immediately following the examination. This option should only be selected where candidates have submitted for examination theses bound in the correct form for final submission (hard-bound, in University of London colour and format). You should then return the corrected copies to the Research Degrees Office along with your reports and this form. **Please also include a list of the amendments made.**

b) You may advise the candidate directly of the necessary amendments and return the thesis copies to them. You should indicate to the candidate (and below) to which examiner they should send an amended copy for checking. When the corrections have been checked you should inform the Research Degrees Office.

c) You may ask the Research Degrees Office to advise the candidate of the necessary amendments. You should return the copies of the thesis to the Research Degrees Office with a list of the amendments you require the candidate to make and indicate below to whom the thesis should be sent for checking that they have been completed satisfactorily.

Please note that it is not permitted for supervisors to be nominated to check amendments.

Please indicate which of the options outline above applies:

a) The candidate has completed the amendments to our satisfaction. The corrected copies and a list of the amendments made are enclosed with this report. ☐

b) The candidate has been provided with a list of amendments and been asked to send the thesis to ..... JOHN BARNDEN ..... for checking that they have been completed satisfactorily ☑

c) The Research Degrees Office is asked to send the attached list of amendments to the candidate. The corrected thesis should be sent to ..... for checking that they have been completed satisfactorily ☐

Signatures

~~Internal~~ External One examiner:	M.A. Barnard	Date:	6 Nov. 17
External examiner: Two	Anna Kwa	Date:	6.11.2017

Please return this form, together with the joint report, preliminary reports, list of amendments (if applicable) and theses (if applicable) to the Research Degrees Office.

Research Degrees Office
GC 213
Graduate Centre
Queen Mary University of London
Mile End Road
London
E1 4NS

+44 (0)20 7882 7810 (fax)
researchdegrees@qmul.ac.uk

Final Joint Report
on PhD examination on 6 November 2017

Candidate: Stephen McGregor

Thesis title: *Geometric Methods for Context Sensitive Distributional Semantics*

Grounds for decision (Pass subject to Minor Amendments):

Both examiners were impressed by the overall quality of the thesis before the viva. They used the viva to obtain some straightforward minor clarifications and to raise certain discussion issues with the intention of probing further the candidate's ability in academic argument, willingness to be self-critical, and research vision for the future. These qualities were already well attested by the thesis, and were confirmed by the viva. We are confident that the candidate easily meets the requirements for award of a PhD, including on originality, knowledge of the field, ability to conduct theoretical and empirical investigation, and self-directedly defining a significant and timely research endeavour.

The issues mentioned above are laid out in the two preliminary reports, which we have indicated should be handed to the candidate.

The minor amendments are being sent direct to the candidate. Apart from the need to correct some typos, they concern the need for occasional clarification of the text and addition of small amounts of commentary.

Examiner 1: John Barnden Signature: J.A. Barnden Date: 13 Nov. 2017.

Examiner 2: Anna Korhonen Signature: Anna Korhonen Date: 13.11.2017

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.

PRELIMINARY REPORT on the PhD Thesis by Stephen McGregor

External Examiner 1: John Barnden

The thesis sets out a novel and well-motivated line of investigation in the area of human language semantics called distributional semantics, with the aim of making distributional semantics flexibly responsive to context and also more theoretically transparent. The thesis provides systematic empirical evidence that the methods proposed can form the basis of automated systems that perform competitively or better on salient tasks being addressed in the distributional semantics area, notably the advanced task of metaphor detection. Most importantly, the theory and experiments conducted allow a novel and revealing type analysis of semantic phenomena, allowing new insights about the powers of and challenges for distributional semantics.

Overall the thesis makes convincing case for the practical and theoretical worth of the candidate's central idea of using geometrically richer mathematical spaces for representing statistical semantic information than those used in other approaches in the area. The candidate also intelligently discusses interactions of the work with major philosophical contentions about the nature of language. The work should stimulate new research by others and should be of interest to a wide academic audience.

I anticipate that the viva will confirm my current impression that the candidate should be awarded a PhD with, at most, minor revisions. I have some issues and discussion points to raise at the viva, but they do not constitute significant objections to the thesis, and I anticipate that the candidate will be able to address them in an intelligent and satisfactory way. Aside from a few clarifications on detail that I would like to receive, the main issues/points are:

-- whether/how the approach addresses certain aspects of metaphor not covered in the thesis

-- the extent to which the approach could be (made) sensitive to very local contexts arising in a particular discourse

-- a need for more insight into why particular, apparently irrelevant dimensions (e.g. "hearing", "accidentally") arise as important in an analogy experiment in Chapter 7

-- whether the claims about metaphor by the philosopher Donald Davidson, cited positively at various points in the thesis, conflate issues of indeterminacy and issues of non-propositionality

-- (speculatively) ways in which the method could possibly be broadened to deal with other channels of communication such as images and gestures.

J.A. Barnden

3 Nov. 2017.

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

Language Technology Lab
Department of Theoretical and Applied Linguistics
English Faculty Building, 9 West Road

Tel: +44 (0) 7710 954346

Email: alk23@cam.ac.uk
http://www.cl.cam.ac.uk/~alk23/