Graduate Research School

Examiners' Recommendation Form
Degree of Doctor of Philosophy

"The (PhD) thesis shall be a substantial and original contribution to scholarship, for example through the discovery of new knowledge, the formulation of theories or the innovative re-interpretation of known data and established ideas"

CANDIDATE:	Mr L White	REF:	20361362
EXAMINER:	ASSOCIATE PROFESSOR J ZHANG		

RECOMMENDATION FOR CLASSIFICATION

- Please mark <u>one</u> box only and refer Page 2 for the required Merit Criteria Scores

PASS	
☑	The thesis be **PASSED** with no requirement for correction or amendments and the student be awarded the degree of Doctor of Philosophy.
☐	The thesis be **PASSED, SUBJECT TO MINOR REVISION** as indicated in my report to the satisfaction of the Board.
☐	The thesis be **PASSED, SUBJECT TO SUBSTANTIVE* AMENDMENTS** along the lines indicated in my report. *The student will normally be instructed to submit a detailed report to the Board of the Graduate Research School outlining the amendments to the thesis. The Board will consider the report and the revisions in determining the final classification, without further reference to the examiners.
(if applicable) ☐	I believe this to be in the top 5% of thesis I have examined.

OR

RE-SUBMIT	
☐	The thesis be returned to me for re-examination after completing the required extra work and revision indicated in my report. A thesis which must be re-submitted requires alterations of such scale, complexity and/or conceptual significance that their adequacy should be appraised again.

OR

AWARD OF THE DEGREE OF MASTER	
☐	The thesis NOT be awarded the degree of Doctor of Philosophy but be PASSED for the appropriate degree of MASTER. (The (Masters) degree must be a substantial work generally based on independent research which shows a sound knowledge of the subject of the research, evidence of the exercise of some independence of thought and ability of expression in clear and concise language).

OR

FAIL	
☐	The thesis be FAILED and the student NOT be awarded the degree of Doctor of Philosophy OR the degree of master and NOT be permitted to resubmit the thesis in a revised form.

-Please also complete and return the following page with the merit criteria scores-

Graduate Research School

Examiners' Recommendation Form
Degree of Doctor of Philosophy

CANDIDATE:	Mr L White	REF:	20361362
EXAMINER:	ASSOCIATE PROFESSOR J ZHANG		

Please provide a Merit Criterion Score for each of the six criteria below, according to the key at the bottom of the page.

SCORE	CRITERIA FOR THE AWARD OF DOCTOR OF PHILOSOPHY - THESIS
2	The thesis as a whole is a substantial and original contribution to knowledge of the subject with which it deals.
1	The student shows familiarity with, and understanding of, the relevant literature.
1	The thesis provides a sufficiently comprehensive study of the topic.
2	The techniques adopted are appropriate to the subject matter and are properly applied.
1	The results are suitably set out, and accompanied by adequate exposition.
1	The quality of English and general presentation are of a standard for publication.
SCORE	**CRITERIA FOR ANY ADDITIONAL EXAMINABLE COMPONENT/S FOR AWARD OF DOCTOR OF PHILOSOPHY –(IF APPLICABLE)**
	Any additional examinable components demonstrate technical accomplishment and imaginative resource and/or advanced technical and interpretative accomplishment (as appropriate)

Please attach your report (minimum length – 1 page).

☑ I confirm that there is no actual or perceived conflict of interest arising from my examination of this thesis.

Signature: _____ Date: 04/12/2018

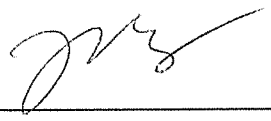
Merit Criteria Score Key

SCORE	MERIT CRITERIA	DESCRIPTION
1	**Exceptional**	Of the highest merit, at the forefront of international PhDs in the field. Fewer than 5% of students worldwide would fall in this band.
2	**Excellent**	Strongly competitive at international levels. Fewer than 20% of students would fall in this band.
3	**Very good**	An interesting, sound and compelling thesis. Approximately 30% of students would fall in this band.
4	**Good**	A sound thesis, but lacks a compelling element in some respect. Approximately 30% of students would fall in this band.
5	**Fair**	The thesis has potential, but requires major revisions. Approximately 20% of students would score in this band.
6	**Flawed**	The thesis does not meet the required standard for this criterion

THE UNIVERSITY OF WESTERN AUSTRALIA

NON-DISCLOSURE STATEMENT BY AN EXAMINER OF A Doctor of Philosophy THESIS

I, Associate Professor J Zhang, acknowledge and accept that the information contained in the thesis entitled "**On the surprising capacity of linear combinations of embeddings for natural language processing**" which has been submitted by **Mr L White** (the author) for the degree of Doctor of Philosophy at The University of Western Australia (the University) is confidential. In consideration of the University appointing me an examiner I undertake not to disclose or discuss any of the information contained in the thesis without the written permission of the author and the University for a period of **twelve months** from the date on which I receive the thesis for examination.

Signature 

Date 04 December 2018

Examiner's report for

PhD thesis

On the surprising capacity of linear combinations of embeddings for natural language processing

Lyndon White

Contribution to knowledge

This thesis presents coherent research on applying linear combination of pre-trained word embeddings to a number of practical natural language processing tasks. It has shown that this simple input representation technique is surprisingly powerful. The utility of this representation is demonstrated in a number of natural language understanding tasks.

The thesis thus makes nice contributions solving some practical NLP tasks using the simple language model of linear combination of pre-trained word embeddings. The research results presented in this thesis also have potential deep impact on the theoretical research on language modelling. The significance of research findings in the thesis is evidenced by several publications. The research results are published as several conference papers, including ACL, a prestigious conference in the NLP field and another paper is in submission to Journal of Computational Linguistics, a top NLP journal. The research results on generating bags of words from sums of their word embeddings are published in an international conference and won the best student paper award.

Beyond the main content for research results, the dissertation also includes an appendix detailing software tools.

Contents

The thesis comprises an introduction, two main parts – Part I is a literature review and Part II is a collection of publications, and a conclusion. Part I of the current literature on neural models for language representation at different levels is self-contained and well presented. The contents are literature review are based on a published book authored by the student, which demonstrates his deep understanding of the current literature on representation learning for language understanding.

Part II is a collection of publications which demonstrate the utility of linear combination of word vectors for several practical NLP tasks, , including sentence paraphrase classification, colour estimation for short phrases and named entity identification, construction of bag of words from sentence representations and further generating sentences.

Writing

The thesis is overall well written. The thesis is well organised and easy to follow.

Compared to the other parts of the thesis, the Conclusion chapter is a bit short. The utility of the proposed simple language model is nicely summarised. Section 11.1.2 discusses that linear combination of word embeddings complements the neural language models, but needs to be strengthened. Some deeper discussions of using the combination of word embeddings as a language model in the context of the current literature on neural language models would better highlight the potential theoretical implications of the research. Furthermore discuss if it is justified to further investigate strategies to encode the signals from the word embedding combinations into advanced neural language models for more accurate language models in general.

There are some minor expression errors that need to be fixed. For example:
… their intimidate parents are their POS tags
… some citation format issues. For example section 5.1: …These tasks include: word prediction Le and Mikolov 2014; recreation of input, as in the auto-encoders of Socher et al. 2011c; Socher et al. 2011a and Iyyer, Boyd-Graber, and Daumé III 2014;

Graduate Research School

Examiners' Recommendation Form
Degree of Doctor of Philosophy

"The (PhD) thesis shall be a substantial and original contribution to scholarship, for example through the discovery of new knowledge, the formulation of theories or the innovative re-interpretation of known data and established ideas"

CANDIDATE:	Mr L White	REF:	20361362
EXAMINER:	PROFESSOR D POWERS		

RECOMMENDATION FOR CLASSIFICATION

- Please mark <u>one</u> box only and refer Page 2 for the required Merit Criteria Scores

PASS	
☐	The thesis be **PASSED** with no requirement for correction or amendments and the student be awarded the degree of Doctor of Philosophy.
☐	The thesis be **PASSED, SUBJECT TO MINOR REVISION** as indicated in my report to the satisfaction of the Board.
☑	The thesis be **PASSED, SUBJECT TO SUBSTANTIVE* AMENDMENTS** along the lines indicated in my report. *The student will normally be instructed to submit a detailed report to the Board of the Graduate Research School outlining the amendments to the thesis. The Board will consider the report and the revisions in determining the final classification, without further reference to the examiners.
(if applicable)	
☐	I believe this to be in the top 5% of thesis I have examined.

OR

RE-SUBMIT	
☐	The thesis be returned to me for re-examination after completing the required extra work and revision indicated in my report. A thesis which must be re-submitted requires alterations of such scale, complexity and/or conceptual significance that their adequacy should be appraised again.

OR

AWARD OF THE DEGREE OF MASTER	
☐	The thesis NOT be awarded the degree of Doctor of Philosophy but be PASSED for the appropriate degree of MASTER. (The (Masters) degree must be a substantial work generally based on independent research which shows a sound knowledge of the subject of the research, evidence of the exercise of some independence of thought and ability of expression in clear and concise language).

OR

FAIL	
☐	The thesis be FAILED and the student NOT be awarded the degree of Doctor of Philosophy OR the degree of master and NOT be permitted to resubmit the thesis in a revised form.

-Please also complete and return the following page with the merit criteria scores-

Graduate Research School

Examiners' Recommendation Form
Degree of Doctor of Philosophy

CANDIDATE:	Mr L White	REF:	20361362
EXAMINER:	PROFESSOR D POWERS		

Please provide a Merit Criterion Score for each of the six criteria below, according to the key at the bottom of the page.

SCORE	CRITERIA FOR THE AWARD OF DOCTOR OF PHILOSOPHY - THESIS
4	The thesis as a whole is a substantial and original contribution to knowledge of the subject with which it deals.
5	The student shows familiarity with, and understanding of, the relevant literature.
4	The thesis provides a sufficiently comprehensive study of the topic.
3	The techniques adopted are appropriate to the subject matter and are properly applied.
4	The results are suitably set out, and accompanied by adequate exposition.
2	The quality of English and general presentation are of a standard for publication.
SCORE	**CRITERIA FOR ANY ADDITIONAL EXAMINABLE COMPONENT/S FOR AWARD OF DOCTOR OF PHILOSOPHY –(IF APPLICABLE)**
	Any additional examinable components demonstrate technical accomplishment and imaginative resource and/or advanced technical and interpretative accomplishment (as appropriate)

Please attach your report (minimum length – 1 page).

☑ I confirm that there is no actual or perceived conflict of interest arising from my examination of this thesis.

Signature: _____ Date:_____

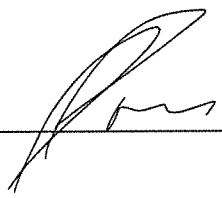
Merit Criteria Score Key

SCORE	MERIT CRITERIA	DESCRIPTION
1	**Exceptional**	Of the highest merit, at the forefront of international PhDs in the field. Fewer than 5% of students worldwide would fall in this band.
2	**Excellent**	Strongly competitive at international levels. Fewer than 20% of students would fall in this band.
3	**Very good**	An interesting, sound and compelling thesis. Approximately 30% of students would fall in this band.
4	**Good**	A sound thesis, but lacks a compelling element in some respect. Approximately 30% of students would fall in this band.
5	**Fair**	The thesis has potential, but requires major revisions. Approximately 20% of students would score in this band.
6	**Flawed**	The thesis does not meet the required standard for this criterion

THE UNIVERSITY OF WESTERN AUSTRALIA

NON-DISCLOSURE STATEMENT BY AN EXAMINER OF A Doctor of Philosophy THESIS

I, Professor D Powers, acknowledge and accept that the information contained in the thesis entitled "**On the surprising capacity of linear combinations of embeddings for natural language processing**" which has been submitted by **Mr L White** (the author) for the degree of Doctor of Philosophy at The University of Western Australia (the University) is confidential. In consideration of the University appointing me an examiner I undertake not to disclose or discuss any of the information contained in the thesis without the written permission of the author and the University for a period of **twelve months** from the date on which I receive the thesis for examination.

Signature 

Date 2 Nov 2018

Lyndon White – PhD Examiner Report

This thesis has two parts, the first is based on a published book (with the first chapter replaced by an Introduction to the thesis), and the second is based on published conference and workshop papers and a submitted journal paper. Rather than providing a separate introduction and conclusion to the paper, or editing the paper into the flow of the thesis, each paper is included relatively unmodified (although includes information about where it was published, presented or submitted). There is however, an extended abstract (1-2 pages) of each paper dumped into the middle of the introduction. The introduction is totally *unacceptable*, as it introduces technical words (like *embedding*) from the first sentence without definition or explanation, and presents extended abstracts of papers before discussing the aims and directions and approach of the thesis.

A further complication is that chapter 1 of the book is omitted but includes detail that the readers is referred to at times. Additionally, the chapters of Part I are then chapters of a published book (Springer monograph, but actually more like a textbook or focused review than a research monograph) and are intended to serve as the primary literature review for the thesis, but will be omitted from the archived version of the thesis. This is *not acceptable*. The thesis needs to stand alone, and all material needed to understand a concept needs to be introduced before being discussed, at least at the level of comparing and contrasting different approaches. It is reasonable to refer to the book for further tutorial discussion of the espoused approaches.

There is also a set of appendices that relate to open software and are in the form of short papers submitted to open software venues.

Neither the thesis nor the book does a thorough review of the antecedents and alternatives to the approaches adopted and adapted, including proper attribution of the origin of ideas to the original researchers by direct citation (or quotation) of the seminal papers. This is also *unacceptable*. Furthermore, there is a *lack of proper presentation of results* in the published and submitted papers and in addition *no significance testing, no use of confidence intervals*. There are also various other errata both in Part I and in the published and submitted papers.

Further information on all these points are given in the notes below. Some additional points that need addressing are highlighted with dot points, while other suggestions are made that the student might follow to improve the presentation.

The issues raised, and particularly the need to rewrite and expand Chapter 1 (or turn into multiple chapters) and actually perform a proper critical literature review, as well as provide for the removal of the book chapters. Generally any chapters that can't be included should be replaced by a précis or else pointed to by the relevant parts of the actual review chapter. The book does *not* provide an adequate review of the emergence and usage of embeddings.

Ch.1. Introduction

The terms *embedding* and *skip-gram* are used in *Introduction* without proper (explicit) definition. Footnote 2 (equations 1.4-> 1.5) - it is actually a weaker assumption than *independence*, relating rather to the Bayesian concept of *conditional independence*. This could do with further explanation/citation - although perhaps we are getting into too much detail too soon, before basic concepts are introduced. Similarly for RNN and RvNN - it does not suffice to give a source, you need a definition (preferably a quote from the source, or else one converted to your own notation/terminology).
In the introduction, it is perhaps better to confine oneself to general aims and direction and defining the scope and intentions of the thesis (which is not particularly clear other than exploration of embeddings).

- Give clear definitions of all *concepts* used (with citation to the original/seminal work behind them).
- Include a table of acronyms/letterisms/buzzwords (or even hyperlinked/popup definitions) as it is hard to keep track of random combinations of letters.

LCOWE - linear combintions of ...
SOWE - sum of ...
MOWE - mean of (normalized over size of BOW)
Skip-gram (implicitly defined by equations 1.1 to 1.5)
RNN vs RvNN (recurrent vs recursive - what is the difference? not defined at all)

Appeal to Grice's maxim is indeed appropriate. However, puns and garden path sentences do occur naturally (accidentally as well as deliberately) and when accidental may be accompanied by a correction (usually a repair that repeats an elaborated version of the phrase) or an apology about the pun (in either case can get an explicit or implicit claim that it was or wasn't intended). In fact, there is a high level of ambiguity in natural language at the sentence level, including the phonetic level, that is only resolved by topic and local linguistic and/or ontological context.

The overview table 1.1 is very useful - and does gives some impression of the scope of the exploration, but the nature and details of the task could be summarized and justified in more detail. Most of them seem very specific, and not really able to get at the issues relating to the massive polysemy and ambiguity of language as well as the role of word order and morphology, although the recovering sentence task (10) promises to give a reasonably fair evaluation here - although it does not test the correct interpretation in terms of syntax, semantics and pragmatics.

The abstracts of the original papers is perhaps more detail than is needed here, and not a substitute for a cogent summary of the goals and methodological distinctives, although the key contribution paragraph helps here. It moreover mitigates *against* the reader getting a good overview of the work and makes it difficult to refer back to table 1.1 when reading the expected overview in 1.2. I would suggest either removing these abstracts or moving them to a separate section 1.3 after the motivation of the work in 1.2 which does help to round out where the thesis is headed.

The justification of the color understanding work (Ch6) in terms of testing the ability to test for distribution of possible meanings is well taken. The same technique (FastText) is also used for other descriptive terms (Ch8: adjectival contexts inPOV

character detection). What is especially strange is that Ch6 is paired with Ch5 in the discussion, with Ch8 noted as atypical while using the same embedding system.

The use of four different embedding techniques across six different tasks reflects a lack of orthogonality and coherence in the methodology and needs to be justified.

The end of Chapter 1 gives me a good idea of the goal, and appropriate one for me and the many researchers who use BOW representations but feel that they are missing out on the deeper levels of understanding of language and the contextual, ontological and cognitive processes that underlie it.

Basically Chapter 1 needs to be substantially edited and reorganized if not rewritten.

Given Chapters 2 to 4 will be omitted from the archived version, and given the lack of historical and critical review it seems Chapter 1 needs to be extended to provide a standalone thesis for someone reasonably familiar with NLP/CL, even with the book chapters omitted (but presumably replaced by links for purchase or doc.delivery). This would mean that Chapter 1 provides an introduction with the aims and directions of the project, and introduction to key concepts (including defining embeddings, bag of words, etc., as well as providing relevant introductions to the NN and Bayesian aspects for which references to Chapter 1 of the Springer book are given. Then we should get the outline mapping out the thesis and its structure/arguments, and finally the list of published papers and extended abstracts if this is needed here.

An alternative would be to provide an additional literature review chapter, and to provide introductory *glue* at the beginning of each chapter to explain how it fits into the structure of the thesis, rather than assuming that they can keep this in mind from the start.

Additionally, a glossorary or table of Acronyms/Letterisms would be helpful - and a good principle is that all such technical terms/abbreviations get spelled out the first time in each chapter (should already be done for incorporated papers).

Ch.2. Word Representations - book chapter (2nd of 4, citing 1st for Bayes form) (appropriately quoting Firth, but don't hyphenate ragged text)

Figure 2.1 - it would be useful/usual to see different dimension pairings - here particular sets of orthogonal dimensions are collapsed fairly arbitrarily. The words "rook" and "castle" are not ones I'd expect to see (and far from each other - rook near the ways of moving group and castle near the places for living/storing - with car at the boundary rather than rook (in the sense castle) which is along way from chicken and duck (in the sense food rather than bird) and turkey (in the sense of country, and near the home words - presumably why included); interesting to see that opposites are nicely paired, as well as near synonyms, and good topicality. Also note that the adjectival words congregate around the y-axis (x=0), and the more action-related (verb or nouns from verbs) words are close to the x-axis (y=0 for x<0)

Errata

- Back-off *apportions* the probability (not *portions*)

- Cite the inventor of the technique (Bayes-like reformulation) not just your own writings (or any reviewers review) on the subject and give it a proper pedigree.

Possibly part of chapter 1 of the book needs to be included for the thesis to stand on its own, or else a proper introduction to each technique used and acknowledgment of the full chain of originators/developers of that technique. This proper treatment is essential for any theory/technique that is a core support of any aspect of the work presented.

- The layer-bypass is a *contrivance* to aid in the learning. (not *connivance*)
- not present in *any/most of the work* deriving from this (not *future works* - *work* is a mass noun except in the sense of *opus*)
- This network is used in the thesis as a basis for explaining the idea of using neural networks with vector representations of words. (they are not vague *notions*, and neural networks using bag of word and Ngram representationsgoes back to at least the 70s, and there was related work with speech even in the 60s, and considerable work in the 80s and 90s).

In fact, there was an entire AAAI symposium on connectionist approaches to natural languages around 1991, as well as featuring strongly in more general events that focussed on learning approaches to language in the early 90s. In fact, already in the 80s there was discussion of different ways of dealing with the context (bag of word/continuous word, unidirectional/bidirectional, supervised/unsupervised, stopping/stemming/both/neither, biologically/computationally motivated) - and at least half a dozen PhD theses by 1995. So saying "Bengio et al. (2003) ... is the network which begins the notions of using neural networks with vector representations of words. Bengio et al. focused on the use of the of sliding window of previous words" is totally wrong. They were around four decades too late to get that credit!

Similarly something like the misnamed CBOW was being used in 1980s (and there was already competition/dissension between those using the continuous approach and the BAG approach by the early 1990s. This is talking NL/CL - quite apart from the usage in IR as early as the 60s.

There was life before the internet... Your review seems to think things started around 2000... The main thing that happened with the internet is that we got bigger datasets...

In terms of what you do go into detail about, the recent work of people like Bengio and Miklov, the treatment is good, but what about other people? This work wasn't done in a vacuum, and these terms (embedding, BOW, etc.) became very commonly used very quickly (although in fact the approaches were really not that new in the broad sense, although the specific techniques and formulations and derivations and demonstrations were important contributions. What about all the relevant competitions, shared tasks etc?

Ch.3 Word Sense Representations (Ch 4 of same Springer book)

This thesis is unusual in that much of it has been published as a book prior to the thesis being examined, without being through the rigours of journal publication/refereeing. Furthermore, Chapter 1 of the book is cited as if the reader should have access to it (though we don't - or didn't - I have now obtained it and material in chapter 1 would be useful background for many readers of the thesis). The book is written more like a textbook than a research monograph, and itself lacks the citations that are the backbone of academic work. It is not clear whether the discussion of the relationship between softmax and Bayes's theorem (Chapter 1 as cited in the thesis) is meant to be original work, but of course Softmax was designed with a full understanding of this framework and what it means in terms of conditional probabities and conditional independence. Although there a few citations in the book text they are extended to a reasonable list of references at the end of each chapter (although there appropriate reference to seminal work, there is little reference to other authors' work that builds on it other than self-citation).

errata (corrections made in italics without further explanation)

Due to the *overwhelming* frequency of the most frequent sense, it is unlikely for even a small training corpus to have the most frequent sense *differ significantly* from the use in the language as a whole. [I'd drop the unlikely bit - in fact it is likely that some small corpora (extracts) will only use a secondary sense but this should not show up as a significant difference more than about alpha proportion of the time (e.g. for alpha =.05), while a larger corpus is likely to allow discovery of significantly different contexts for most sense eventually (those where a whole raft of language is imported into different contexts along with a consistent metaphor, may not be distinguished - *literally* **1a/b/c** or **2a/b** or **3a/b** subsenses would be hard to distinguish (indeed **1c** is people misusing the word as an intensifier because they haven't caught the proper meaning, only its intensity correlate), although the three primary senses would be relatively easily disambiguated by context.

Generally the discussion here is good, with good critique/intuitions re the work discussed (although still very selective/representative). But statistical WSD techniques (from 15+ years ago before the shift to NN/DNN/LSTM/embedding approaches did very well with these problems, with basic clustering WSD/WSI techniques doing nicely even 30+ years ago). I do note some references to broader than NN approaches back to 1998/2002.

Ch4 Sentence+ Representations (Ch5 of book)

- Webster definition of sentence begs some questions

Is a question a sentence - particularly when part is missing as unknown?
Is a sentence with elided words - from immediate spoken or other context?

Note that in the above two cases part of the complete idea is provided by the other speaker (before or after).

The initial work ... was on the creation of parsers. Again parsers have been around decades before the work cited, and no real introduction to parsing is given in this section (4.33 Parsing). Hinge loss should also be explained properly (with equation).

For Recursive Autoencoders, again refers to Chapter 1 of the Springer book - why not just include this (or parts thereof) - it could fit very nicely into the framework of chapter 1 of the thesis with some sections dealing with assumed knowledge/basics.

Again recursive and expanding/unfolding networks (and autoencoders) go back to at least the 1980s.

This ends Part I - basically the Springer book extract and broad lit.review (which will NOT be included in the archived public version of the thesis) forcing people to buy the book for the expected review part of the thesis. Fortunately, because Part II is a copy on full published refereed papers, they will all have their own appropriate lit.review, with prompting from referees where necessary to add anything missing.

Note that the Open Software papers are all under review (perhaps should have been put up on arXiv to allow prepub access) - these papers and associated software are important (shared) outputs of the thesis (appendices). The CL paper/chapter on colours is also under review

Ch5 (Meaning - Aus.Doc.Comp.Symp.)
Provides a paragraph sketching each component of the background to the work. There is mention of early relevant work back to the 60s.

errata

synonymity -> synonymy

as well as the stop-words: "and", "a" and "of".
were these the *only* stop-words skipped as implied? or did you mean
as well as stop-words like "and", "a" and "of".
Should give a count for each category (numerals, punct, symbols, prop, unusual, stop).

There are mangled refs whether the authors are given twice (without brackets around the date as should generally be there in Harvard style) - e.g. Yin and Schütze 2015, Ritter et. al. (doubling occurs twice - no dot after *et* either) and Le and Mikolov 2014

Weak conclusion about summing being surprisingly effective.

Ch6 (Learning Colours - CL *submitted*)
Should include some discussion of the psychophysiology of colour, as well as the different colour spaces, and in particular the issues with HSV, HSL etc. and the more bioplausible nature of YUV, YCbCr, CIELab/L*a*b* as well as the sociological conditioning and the effects of time of day, background lighting and reflected light, and surround contrast, etc.

Why do you use FastText? Why don't you compare other approaches?

Conditional Independence has particular problems in this domain because the particular way a colour is named (even potentially the same colour described differently) may bias the distribution of acceptability in terms of hue, saturation and value (which is the wrong system to use, as demonstrated repeatedly over the last century or so, and particularly the last decade or so). The problem with HSV can be illustrated by the (mentioned) fact that in HSV people/animals/skin are basically red which means H is unstable at a flip point at 0°=360° (or 0=256 mod 256 for an 8-bit representation, or [0,1) for a real representation) - it is not used in cameras or video or printing, but is arguably intuitive for colour chooser GUI purposes (although

different software companies use different HS* variants). But there are many other problems. It also involves expensive non-linear trig functions.

- So it is important to tabulate the 16 color-spaces evaluated.

The obvious space to use would be L*a*b* which is designed to match/map human perception, although YUV/YCbCr would be what I would probably use (and is the basic of MPEG/JPEG compression). Possibly the high conditional independence for HSV is due to its lack of bioplausibility and relates to the difficulty of calibrating in this system. Another factor is that the names used are combinatiorial (like purple-pink or bluish grey, but not clear why not pinkish purple or purplish pink) rather than reflecting natural semantic/ontologic categories like bluegrass, forest green, or the various military camoflage schemes and khaki variants. This is only testing examples which one would expect would be on a fairly direct path between the pivot colours included in the name (in whatever colour space was used).

There are also lower bit representations versions of YUV etc. that are commonly used.

Note that normalizing to 0-1 for the color spaces is quite a standard operation (and is for example assumed in Matlab toolboxes if using reals rather than ints). The compensation for the hue wrapping issue proposed is also standard, but no citations are given.

For me the weak points of this chapter was failing to use a bioplausible colour space (or justify quantitatively the choice of one of the most difficult to work with) and failing to use a sociologically/culturally selected/conditioned colour lexicon. This makes it hard to know how to interpret the results meaningfully.

There is no indication of significance or confidence, and no standard errors (or confidence intervals or credible intervals), no ROC graphs, no indication of how stable these behaviours are.

Ch7 - Known Meanings (Conf. on Int. Text Proc. and Comp. Ling., 2018)

This paper introduces what it needs well (in terms of a conference paper).

Use of AdaGram is not really justified or explained, although it is briefly characterized.

Again there is no indication of significance or confidence, and no standard errors (or confidence intervals or credible intervals), no ROC graphs, no indication of how stable these behaviours are. Remember 99% word-level accuracy means most of your sentences will have errors (on average over natural text). Remember 0.797 precision from 79.97% attempted is less than 0.637 if unattempted questions are marked wrong (matching Recall for your Refitted-S Greedy in table 7.3).

errata

Note that in several places (caption of table 7.3 and sentences above and below it you use MSF when I presume you mean MFS). Smoothing is a more principled approach than backoff to MFS and your results seem to justify this (*seem* because no s.e.s, C.I.s or p-vals). It would be normal to mark results that are significantly better than baseline

and significantly worse than your proposed method - or else put on standard error information at let people draw their own significance/confidence conclusions (the modern preferred approach called "the new statistics") or if you prefer the Bayesian Credible Intervals, calculate those.

Ch8 - NovelPerspective - PoV Characters (ACL Demo)

This is a nice approach and application.

The issue of exotic names is interesting - one I've been facing in reverse, trying to allow sensible modeling of what characters people might type/mean including when they don't obey the phonetic/orthographic rules of English (as in a company name like IBM).

again no significance/confidence

Ch 9 - BoW from SoWE - Conference on Intelligent Text Processing and Computational Linguistics, in 2016 - Best Student paper

Resynthesis task - very important contribution, although here only to BOW level. Note that some models are intrinsically reversible (or have reversed counterparts).

GloVe used here - again with a vague claim that many embeddings will work (vague in the sense of unsubstantiated, no pilot results presented to support this - should include the results from Skipgram - even as a chapter or online appendix if you really can't fit it - in this case can add as an appendix).

It would be interesting to see Fig 9.2 with reciprocal sentence length (or sentence rank according to Zipf's Law) which would avoid the clumping at the left end. A logarithmic plot would give more detail about the low Jaccard as well as the small sentences. Why Jaccard?

Ch10 - Sent from SoWE - HDDM Workshop (IEEE IntConf on DM) - interesting venue for it

Builds on BoW (Ch9)

LSTM is mentioned here again - and has been a surprising omission from the review. Ordering is based on a simple trigram model (unreported results from beam search mentioned - another appendix that I'd like to see, but at least you mentioned results and significance).

Nice to see appendix with examples of how it works.

Conclusion/Future Work

This just stopped... Note it is conclusions in the sense of thinking about a problem and drawing multiple conclusions (plural) and correctly future work (singular as mass noun).

You provided one sentence summaries of the chapters but didn't really draw conclusions that would guide people in the future, at least until the future work

section where you do discuss limitations (a starting point as much as a conclusion in terms of being related to model limitations - and not that you provided additional trigram info in Ch10).

I note your (continued) emphasis on adversarial examples. More appropriate is perhaps just real world examples and not being satisfied by just getting the short sentences right.

APPENDICES

Good on you for creating/sharing useful tools...

Good on you for critiquing your own work and using it to motivate the toolset...

GENERAL

World without Web

The entirety of this thesis (and even more particularly the book) is based on certain key work relevant to the thesis while ignoring the huge relevant pre-www literature. It not only doesn't even attempt to review the precursors of the main approaches highlighted, it does not trace through the impact and interactions and successors to them.

Thesis without LitRev

Even more problematic is that the thesis is proposed to be archived and diseminated without the review chapters (2 to 4 coming from a published book) that comprise Part I (with thesis Chapter 1 being very poor and prempting many undefined concepts, while including pagelong abstracts of each of the chapters before actually outlining what the thesis is about, its aims, methodology and directions if not conclusions; with book Chapter 1 containing information on Neural Networks and Bayesian interpretations that is referred to in the chapters copied to the thesis, but not adequately dealt with in the thesis itself).

Book without LitRev

If I were reviewing the book, I would be making similar comments that it is important to give a bigger historical context and critique for the evolution of the ideas/approaches presented - given the book is actually well written I'd be inclined to conditionally accept it and suggest adding extra boxes or a final section (as many text books do) to provide this broader (before and after) historical perspective and appropriate attribution.

What do you mean without LitRev?

A literature review should be critical - it should emphasizes what the contribution is, the strengths of the approach and what problem is being overcome by what novel innovations, the weaknesses including the (often hidden) assumptions (or sometimes they are lost in obscurity because of people referring only to textbooks and broad

brush reviews), the difference between optimizing the model and solving the problem, etc. Then it should address what it particularly offers to the problem/aims you have set for yourself, what would be advantages in using it, what would be disadvantages in using it, what would be disadvantages in *not* using it. (Of course it can group and cross group work, and then tabulate it, to emphasize different combinations of ideas that might be used/usable together.)

Papers with LitRev

There is a tendency to refer to other work without really explaining what the distintive contribution was (how it works and what its effect is/was). The refereed papers, whilst they are self-contained, still have (partly due to space limitations) similar issues with the scope and depth of the review - it is important to identify seminal and culminal work for any idea that is discussed, not just refer to one paper that invented a nice name for some combination of existing ideas.

When is a thesis not a thesis?

In all cases I have commented on chapters as such, irrespective of their prior publication. It is important for there to be a definitive readable version of the thesis that stands alone, and editing the papers may be desirable (edited postprints can always be made available too on the archive servers) - and of course, the submitted papers may benefit from examiners comments.

How to fix the thesis?

In this case I would suggest the main thing that needs doing is restructuring chapter 1 to provide the basic outline of the aims and methods of the thesis, go immediately into introducing ideas and hence defining terms, also providing some of the basic details from book chapter 1 and providing proper (forward and backward) historical critique of those ideas. By all means, mention the methods focussed on in chapters 2 to 4 only briefly in this chapter, and note that chapters from the book will provide further tutorial detail, and included in the examiners' version of the thesis, but not in the public version (with link to the book).

The changes to the papers may be made retrospectively (postprints) or just in the thesis, and of course with the CL paper it is likely that there will be some back and forth with referees and examiners comments may reinforce or help here (could potentially provide an update before reviews are back).

Graduate Research School

Examiners' Recommendation Form
Degree of Doctor of Philosophy

"The (PhD) thesis shall be a substantial and original contribution to scholarship, for example through the discovery of new knowledge, the formulation of theories or the innovative re-interpretation of known data and established ideas"

CANDIDATE:	Mr L White	REF:	20361362
EXAMINER:	DR P STENETORP		

RECOMMENDATION FOR CLASSIFICATION

- Please mark <u>one</u> box only and refer Page 2 for the required Merit Criteria Scores

PASS	
☐	The thesis be **PASSED** with no requirement for correction or amendments and the student be awarded the degree of Doctor of Philosophy.
☑	The thesis be **PASSED, SUBJECT TO MINOR REVISION** as indicated in my report to the satisfaction of the Board.
☐	The thesis be **PASSED, SUBJECT TO SUBSTANTIVE* AMENDMENTS** along the lines indicated in my report. *The student will normally be instructed to submit a detailed report to the Board of the Graduate Research School outlining the amendments to the thesis. The Board will consider the report and the revisions in determining the final classification, without further reference to the examiners.
(if applicable)	
☐	I believe this to be in the top 5% of thesis I have examined.

OR

RE-SUBMIT	
☐	The thesis be returned to me for re-examination after completing the required extra work and revision indicated in my report. A thesis which must be re-submitted requires alterations of such scale, complexity and/or conceptual significance that their adequacy should be appraised again.

OR

AWARD OF THE DEGREE OF MASTER	
☐	The thesis NOT be awarded the degree of Doctor of Philosophy but be PASSED for the appropriate degree of MASTER. (The (Masters) degree must be a substantial work generally based on independent research which shows a sound knowledge of the subject of the research, evidence of the exercise of some independence of thought and ability of expression in clear and concise language).

OR

FAIL	
☐	The thesis be FAILED and the student NOT be awarded the degree of Doctor of Philosophy OR the degree of master and NOT be permitted to resubmit the thesis in a revised form.

-Please also complete and return the following page with the merit criteria scores-

Graduate Research School

Examiners' Recommendation Form
Degree of Doctor of Philosophy

CANDIDATE:	Mr L White	REF:	20361362
EXAMINER:	DR P STENETORP		

Please provide a Merit Criterion Score for each of the six criteria below, according to the key at the bottom of the page.

SCORE	CRITERIA FOR THE AWARD OF DOCTOR OF PHILOSOPHY - THESIS
3	The thesis as a whole is a substantial and original contribution to knowledge of the subject with which it deals.
2	The student shows familiarity with, and understanding of, the relevant literature.
3	The thesis provides a sufficiently comprehensive study of the topic.
4	The techniques adopted are appropriate to the subject matter and are properly applied.
3	The results are suitably set out, and accompanied by adequate exposition.
3	The quality of English and general presentation are of a standard for publication.
SCORE	**CRITERIA FOR ANY ADDITIONAL EXAMINABLE COMPONENT/S FOR AWARD OF DOCTOR OF PHILOSOPHY –(IF APPLICABLE)**
3	Any additional examinable components demonstrate technical accomplishment and imaginative resource and/or advanced technical and interpretative accomplishment (as appropriate)

Please attach your report (minimum length – 1 page).

☑ I confirm that there is no actual or perceived conflict of interest arising from my examination of this thesis.

Signature: Pontus Stenetorp Date: 06-12-2018

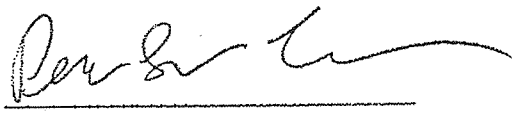
Merit Criteria Score Key

SCORE	MERIT CRITERIA	DESCRIPTION
1	Exceptional	Of the highest merit, at the forefront of international PhDs in the field. Fewer than 5% of students worldwide would fall in this band.
2	Excellent	Strongly competitive at international levels. Fewer than 20% of students would fall in this band.
3	Very good	An interesting, sound and compelling thesis. Approximately 30% of students would fall in this band.
4	Good	A sound thesis, but lacks a compelling element in some respect. Approximately 30% of students would fall in this band.
5	Fair	The thesis has potential, but requires major revisions. Approximately 20% of students would score in this band.
6	Flawed	The thesis does not meet the required standard for this criterion

THE UNIVERSITY OF WESTERN AUSTRALIA

NON-DISCLOSURE STATEMENT BY AN EXAMINER OF A Doctor of Philosophy THESIS

I, Doctor P Stenetorp, acknowledge and accept that the information contained in the thesis entitled "**On the surprising capacity of linear combinations of embeddings for natural language processing**" which has been submitted by **Mr L White** (the author) for the degree of Doctor of Philosophy at The University of Western Australia (the University) is confidential. In consideration of the University appointing me an examiner I undertake not to disclose or discuss any of the information contained in the thesis without the written permission of the author and the University for a period of **twelve months** from the date on which I receive the thesis for examination.

Signature 

Date 10-12-2018

Report on the PhD thesis of Lyndon White

1 General comments

I find this thesis to be sufficient to fulfil the requirements necessary to award a PhD. The author shows a clear understanding of the literature and positions the work and the hypothesis very well relative to ongoing research. Given our current poor understanding of "deep" models, work such as this is very welcome and scratches surface as to why bag of words and "bag of embeddings" models remain competitive in our field. My main objection to the work in this thesis would be that it frequently leaves avenues open for alternative hypotheses, leading to mostly indicative evidence being presented. However, regardless of this, it is clear that scientific contributions – empirical and theoretical – contained within this thesis are sufficient to award the degree.

For comments on a chapter-by-chapter basis, please see my detailed comments below.

2 Point-by-point comments

2.1 Abstract

1. Word embeddings are not explained – even on a high level – in the abstract, this should probably be addressed. In all other respects, the abstract is excellent.

2.2 Chapter 1

1. The introduction is insightful and reads very well, drawing upon Grice's cooperative principle is interesting indeed. The analogies used are appropriate and helps the reader throughout the chapter.
2. In the introduction it is stated that a sum of embeddings has a single hidden layer, according to the common definition it has no hidden layer and a single joint input/output layer.

2.3 Chapter 2

1. In the abstract of Chapter 2 it is stated that word embeddings is what brought machine learning to the forefront of NLP, this is historically dubious given that even prior to this "revolution" statistical methods ruled the roost. Rather, it was the first success of "deep" systems that arguably started the current trend of neural methods.
2. One could cite neural skip-layer connections where it is relevant in relation to Bengio et al. (2003).
3. "Skip-gram is much more commonly used than CBOW.", why?
4. The cosine similarity does indeed ignore the magnitude of the vectors, but it is never explained why this is useful. It is due to the fact that the magnitude is closely related to word frequency, something that has been noted in numerous publications.
5. There are numerous misspellings and grammatical oddities in Chapter 2 that should be addressed, for example "...in the with in the range...".
6. I assume that placeholders will be used in the archival copy of the thesis for the chapters that will not be released?

7. Overall, Chapter 2 presents a very clear picture of the underpinnings of word representations and only have the minor issues noted above.

2.4 Chapter 3

1. Chapter 3 has similar grammatical and spelling issues as Chapter 2, for example "maybe" where "may be" should be used. Other than this Chapter 3 is very clear and a fair summarisation of the current research.

2.5 Chapter 4

1. Yet again in Chapter 4 there are grammar and spelling issues, these should be addressed.
2. Chapter 4 should probably also discuss the work of Conneau et al. (2017) to learn sentence encoders, other than that the chapter is complete and an excellent summary of the literature.

2.6 Chapter 5

1. Odd citation patterns in Chapter 5 without parentheses, fix these.
2. Just like in previous chapters, Chapter 5 contains spelling errors and grammatical oddities.
3. Other than the above comments, the chapter reads well, the results are convincing, the novel evaluation method is well motivated. In hindsight, it is interesting to consider this work in relation to Conneau et al. (2017) that achieved excellent results by training sentence encoders in a supervised fashion on entailment data.

2.7 Chapter 6

1. This chapter, unlike the ones preceding it, has no grammar or spelling issues.
2. While I agree that this work is different from Kawakami et al. (2016), I am not sure as to why the authors could not simply have added a character-based RNN to make their results at least somewhat comparable?
3. Given that a lot of the results could be attributed to overfitting, it would be helpful if the authors stated exactly how the performed model searches so as to not greatly overfit the data with the more "powerful" CNN and RNN models.
4. Overall, I find the work solid and interesting. However, I am somewhat concerned to read too much into what the experiments say about composition with only 76 samples that are word-order dependent.

2.8 Chapter 7

1. As in several other chapters, numerous citations appear odd – perhaps due to changing between LaTeX styles?
2. Other than the citation style oddities, the chapter reads well and the contribution is valid. However, the empirical results are indeed somewhat modest.

2.9 Chapter 8

1. No formatting, spelling, or grammatical errors in this chapter – it was all very clear and it reads well.
2. The contributions of this chapter may be clearly presented and while mainly technical, valid. But they are only loosely attached to the overall thesis of this work and one can wonder if the chapter could not have been left out.

2.10 Chapter 9

1. Something that strikes me when reading the chapter is that the setting is effectively that of an autoencoder, yet, not parallels are drawn to work such as Socher et al. (2011). I also seem to remember work on Restricted Boltzmann Machines to recover discrete bag-of-words features, but the exact work eludes me. Arguably, this work – which pre dates that of this chapter – should have been discussed and compared experimentally.

2. While reading the chapter, it is not clear to me why no comparison is made to a machine learning method. Not that there is anything wrong with the given approach – in fact, I find it very interesting – but the case for it would have been stronger by showing the difficulty of learning the same task.
3. This is one of the most interesting chapters in the thesis, with both theoretical and experimental contributions. It is presented clearly and well argued for, my only objection is that there should be parallels drawn to autoencoder as I outlined above.

2.11 Chapter 10

1. The chapter is clearly written and I have no complaints in terms of the writing.
2. I am again somewhat puzzled as to why no comparison is made to alternative machine learning-based methods or autoencoders.
3. Can the computational limitation of sentences of a length of at most 18 tokens ever be overcome? If so, how?

2.12 Chapter 11

1. The conclusions laid forth are in line with the observations made throughout the thesis. It would however had been welcome if there was additional analysis or references regarding the hypothesis that most sentences are sufficiently unique to be represented using a LCOWE.

2.13 Appendix

1. As the tooling contributions are outside of the main scope of the thesis I defer further comments and simply compliment that the student and his institution take software contributions as real contributions towards a thesis. This is a welcome shift in culture.