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

Lyndon White
BCM in Computation and Pure Mathematics;
BE in Electrical and Electronic Engineering
September 25, 2018

This thesis is presented for the degree of
Doctor of Philosophy
of The University of Western Australia

Thesis Declaration

I, Lyndon White, certify that:

This thesis has been substantially accomplished during enrolment in the degree.

This thesis does not contain material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution.

No part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of The University of Western Australia and where applicable, any partner institution responsible for the joint-award of this degree.

This thesis does not contain any material previously published or written by another person, except where due reference has been made in the text.

The work(s) are not in any way a violation or infringement of any copyright, trademark, patent, or other rights whatsoever of any person.

The work described in this thesis was partially funded by Australian Research Council grants DP150102405 and LP110100050; and by a Australian Government Research Training Program (RTP) Scholarship.

This thesis contains published work and/or work prepared for publication, some of which has been co-authored.

n here] Signature:

Date: September 25, 2018

In memoriam of
Laurie White
1927–2018

Abstract

As Webster's classic 1900 text *English: Composition and Literature* states: "A sentence is a group of words expressing a complete thought." People use natural language to represent thoughts. Thus the representation of natural language, in turn, is of fundamental importance in the field of artificial intelligence. Natural language understanding is a research area which revolves around how to represent text in a form that an algorithm can manipulate in such a way as to mimic the ability of a human to truly understand the text's meaning. In this dissertation, we aim to extend the practical reach of this area, by exploring a commonly overlooked method for natural language representation: linear combinations (i.e. weighted sums) of embedded representations. This dissertation is organised as a collection of research publications: with our novel contributions published in conference proceedings or journals; and with a comprehensive literature review published as part of a book.

When considering how to represent English input into a natural language processing system, a common response is to view it as a sequential modelling problem: as a discrete time-series of words. A more complex alternative is to base the input model on the grammatical tree structures used by linguists. But there are also simpler models: systems based on just summing the word embeddings. On a variety of tasks, these work very well – often better than the more complex models. This dissertation examines these linear combinations of embeddings for natural language understanding tasks.

In brief, it is found that a sum of embeddings is a particularly effective dimensionality-reduced representation of a bag of words. The dimensionality reduction is carried out at the word level via the implicit matrix factorization on the collocation probability matrix. It thus captures into the dense word embeddings the key features of lexical semantics: words that occur in similar contexts have similar meanings. We find that summing these representations of words gives us a very useful representation of structures built upon words.

A limitation of the sum of embedding representation is that it is unable to represent word order. This representation does not capture any order related information; unlike for example a recurrent neural network. Recurrent neural networks, and other more complex models, are out performed by sums of embeddings in tasks where word order is not highly significant. It is found that even in tasks where word order does matter to an extent, the improved training capacity of the simpler model still means that it performs better than more complex models. This limitation thus impacts surprisingly little.

Acknowledgements

TODO

Authorship declaration

Publications

This thesis contains work that has been published and/or prepared for publication.

Details of the work:
Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun (2018a). *Neural Representations of Natural Language*. Studies in Computational Intelligence (Book). Springer Singapore. ISBN: 9789811300615
Location in thesis: ??
Student contribution to work:
Determined content. Created figures. Wrote book. Supervisors reviewed and provided useful feedback for improvement.

Details of the work:
Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun (2015). "How Well Sentence Embeddings Capture Meaning". In: *Proceedings of the 20th Australasian Document Computing Symposium*. ADCS '15. Parramatta, NSW, Australia: ACM, 9:1–9:8. ISBN: 978-1-4503-4040-3. DOI: 10.1145/2838931.2838932
Location in thesis: ??
Student contribution to work:
Devised problem. Designed and implemented algorithms. Conducted experiments. Created figures. Wrote publication. Supervisors reviewed and provided useful feedback for improvement.

Details of the work:
Lyndon. White, Roberto. Togneri, Wei. Liu, and Mohammed Bennamoun (2018). "Learning of Colors from Color Names: Distribution and Point Estimation". In: *Computational Lingustics (Under Review)*
Location in thesis: ??
Student contribution to work:
Devised problem. Designed and implemented algorithms. Conducted experiments. Created figures. Wrote publication. Supervisors reviewed and provided useful feedback for improvement.

Details of the work:
Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun (2018b). "Finding Word Sense Embeddings Of Known Meaning". In: *19th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*
Location in thesis: ??
Student contribution to work:
Devised problem. Designed and implemented algorithms. Conducted experiments. Created figures. Wrote publication. Supervisors reviewed and provided useful feedback for improvement.

Details of the work:
Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun (2018b). "NovelPerspective: Identifying Point of View Characters". In: *Proceedings of ACL 2018, System Demonstrations*. Association for Computational Linguistics
Location in thesis: ??
Student contribution to work:
Devised problem. Designed and implemented algorithms. Conducted experiments. Created figures. Wrote publication. Supervisors reviewed and provided useful feedback for improvement.

Details of the work:
Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun (2016a). "Generating Bags of Words from the Sums of their Word Embeddings". In: *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*
Location in thesis: ??
Student contribution to work:
Devised problem. Designed and implemented algorithms. Conducted experiments. Created figures. Wrote publication. Supervisors reviewed and provided useful feedback for improvement.

Details of the work:
Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun (2016b). "Modelling Sentence Generation from Sum of Word Embedding Vectors as a Mixed Integer Programming Problem". In: *IEEE International Conference on Data Mining: High Dimensional Data Mining Workshop (ICDM: HDM)*. DOI: 10.1109/ICDMW.2016.0113
Location in thesis: ??
Student contribution to work:
Devised problem. Designed and implemented algorithms. Conducted experiments. Created figures. Wrote publication. Supervisors reviewed and provided useful feedback for improvement.

Details of the work:
Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun (2018a). "DataDeps.jl: Repeatable Data Setup for Reproducible Data Science". In: *Journal of Open Research Software (Under Review)*
Location in thesis: ??
Student contribution to work:
Primary author of software. Created figures. Wrote publication. Supervisors reviewed and provided useful feedback for improvement.

Details of the work:
Lyndon White and Sebastin Santy (2018). "DataDepsGenerators.jl: making reusing data easy by automatically generating DataDeps.jl registration code". In: *Journal of Open Source Software (Under Review)*
Location in thesis: ??
Student contribution to work:
Original author of software. Provided direction, guidance, and code review for its enhancement. Wrote publication.

Details of the work:
Lyndon White and David Ellison (2018). "Embeddings.jl: easy access to pretrained word embeddings from Julia". In: *Journal of Open Source Software (Under Review)*
Location in thesis: ??
Student contribution to work:
Original and primary author of software. Wrote publication.

Details of the work:
Lyndon White and Jonathan Malmaud (2018). "TensorFlow.jl: An Idiomatic Julia Front End for TensorFlow". In: *Journal of Open Source Software (Under Review)*
Location in thesis: ??
Student contribution to work:
Co-maintainer and second highest contributor to the software. Co-wrote publication.

Permission to use work in thesis

The coauthors signing below give permission to use the aforementioned works in this dissertation, and certify that the student's statements regarding their contribution to the respective co-authored works listed above are correct.

Roberto Togneri:
Primary Supervisor
todo
dd/mm/yy

Wei Liu:
Supervisor
todo
dd/mm/yy

Mohammed Bennamoun:
Supervisor
todo
dd/mm/yy

Sebastin Santy:

24/09/18

David Ellison:
todo
dd/mm/yy

Jonathan Malmaud:

24/09/18