Annual Report 2017-2018

Lyndon White

March 2, 2018

1 Summary of Research Progress to Date (including any change in focus, and list of publications)

The primary deviation from planning was the request to produce a book for Springer Briefs. This did however overlap significantly with the literature review refining that was already planned as a major task.

1.1 Publications Arising

1.1.1 Accepted

(Conference Paper) Sentences Meaning Capture published ADCS 2015. Will be in Part C (advantages of one or the other) of thesis. Bibliographic details

Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. "How Well Sentence Embeddings Capture Meaning". In: *Proceedings of the 20th Australasian Document Computing Symposium*. ADCS '15. Parramatta, NSW, Australia: ACM, 2015, 9:1–9:8. ISBN: 978-1-4503-4040-3. DOI: 10.1145/2838931.2838932. URL: http://doi.acm.org/10.1145/2838931.2838932.

Abstract: *Several approaches for embedding a sentence into a vector space have been developed. However, it is unclear to what extent the sentence's position in the vector space reflects its semantic meaning, rather than other factors such as syntactic structure. Depending on the model used for the embeddings this will vary – different models are suited for different down-stream applications. For applications such as machine translation and automated summarization, it is highly desirable to have semantic meaning encoded in the embedding. We consider this to be the quality of semantic localization for the model – how well the sentences' meanings coincides with their embedding's position in vector space. Currently the semantic localization is assessed indirectly through practical benchmarks for specific applications. In this paper, we ground the semantic localization problem through a semantic classification task. The task is to classify sentences according to their meaning. A SVM with a linear kernel is used to perform the classification using the sentence vectors as its input. The sentences from subsets of two corpora, the Microsoft Research Paraphrase corpus and the Opinosis corpus, were partitioned according to their semantic equivalence. These partitions give the target classes for the classification task. Several existing models, including URAE, PV–DM and PV–DBOW, were assessed against a bag of words benchmark*

(Conference Paper) SOWE2BOW published CICLing 2016. Will be part of Part B (bridging representations) of thesis Bibliographic details

Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. "Generating Bags of Words from the Sums of their Word Embeddings". In: *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. 2016.

Abstract: *Many methods have been proposed to generate sentence vector representations, such as recursive neural networks, latent distributed memory models, and the simple sum of word embeddings (SOWE). However, very few methods demonstrate the ability to reverse the process – recovering sentences from sentence embeddings. Amongst the many sentence embeddings, SOWE has been shown to maintain semantic meaning, so in this paper we introduce a method for moving from the SOWE representations back to the bag of words (BOW) for the original sentences. This is a part way step towards recovering the whole sentence and has useful theoretical and practical applications of its own. This is done using a greedy algorithm to convert the vector to a bag of words. To our knowledge this is the first such work. It demonstrates qualitatively the ability to recreate the words from a large corpus based on its sentence embeddings. As well as practical applications for allowing classical information retrieval methods to be*

combined with more recent methods using the sums of word embeddings, the success of this method has theoretical implications on the degree of information maintained by the sum of embeddings representation. This lends some credence to the consideration of the SOWE as a dimensionality reduced, and meaning enhanced, data manifold for the bag of words.

(Conference Paper) SOWE2Sent published ICDM HDD Workshop 2016. Will be part of Part B (bridging representations) of thesis. Bibliographic details

Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. "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)*. 2016. DOI: 10.1109/ICDMW.2016.0113. URL: http://white.ucc.asn.au/publications/White2016SOWE2Sent.pdf.

Abstract: *Converting a sentence to a meaningful vector representation has uses in many NLP tasks, however very few methods allow that representation to be restored to a human readable sentence. Being able to generate sentences from the vector representations demonstrates the level of information maintained by the embedding representation ĂćâĆňâĂIJ in this case a simple sum of word embeddings. We introduce such a method for moving from this vector representation back to the original sentences. This is done using a two stage process; first a greedy algorithm is utilised to convert the vector to a bag of words, and second a simple probabilistic language model is used to order the words to get back the sentence. To the best of our knowledge this is the first work to demonstrate quantitatively the ability to reproduce text from a large corpus based directly on its sentence embeddings*

(Conference Paper) Word-Sense Alignment CICLing 2018. Will be in Part B (bridging representations) of thesis. Bibliographic details

Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. "Finding Word Sense Embeddings Of Known Meaning". In: *19th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing* (2018).

Abstract: *Word sense embeddings are vector representations of polysemous words – words with multiple meanings. These induced sense embeddings, however, do not necessarily correspond to any dictionary senses of the word. This limits their applicability in traditional semantic-orientated tasks such as lexical word sense disambiguation. To overcome this, we propose a method to find new sense embeddings of known meaning. We term this method refitting, as the new embedding is fitted to model the meaning of a target word in the example sentence. This is accomplished using the probabilities of the existing induced sense embeddings, as well as their vector values. Our contributions are threefold: (1) The refitting method to find the new sense embeddings; (2) a novel smoothing technique, for use with the refitting method; and (3) a new similarity measure for words in context, defined by using the refitted sense embeddings. We show how our techniques improve the performance of the Adaptive Skip-Gram sense embeddings for word similarly evaluation; and how they allow the embeddings to be used for lexical word sense disambiguation – which was not possible using the induced sense embeddings.*

1.1.2 Under Review

(Conference Paper) ColorDist under-review for ACL 2018. Will be part of Part C (advantages of one or the other) of thesis. Preprint bibliographic details

L. White, R. Togneri, W. Liu, and M. Bennamoun. "Learning Distributions of Meant Color". In: *ArXiv e-prints* (Sept. 2017). arXiv: 1709.09360 [cs.CL]. URL: https://arxiv.org/abs/1709.09360.

Abstract: *When a speaker says the name of a color, the color that they picture is not necessarily the same as the listener imagines. Color is a grounded semantic task, but that grounding is not a mapping of a single word (or phrase) to a single point in color-space. Proper understanding of color language requires the capacity to map a sequence of words to a probability distribution in color-space. A distribution is required as there is no clear agreement between people as to what a particular color describes – different people have a different idea of what it means to be "very dark orange". We propose a novel GRU-based model to handle this case. Learning how each word in a color name contributes to the color described, allows for knowledge sharing between uses of the words in different color names. This knowledge sharing significantly improves predicative capacity for color names with sparse training data. The extreme case of this challenge in data sparsity is for color names without any direct training data. Our model is able to predict reasonable distributions for these cases, as evaluated on a held-out dataset consisting only of such terms.*

1.1.3 Pending

(Book): Neural Representations of Natural Language (NRoNL) In final stage of editing. To be published as a SpringerBrief. 3 chapters of it will form Part A (literature review) of thesis. Bibliographic details
Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. *Neural Representations of Natural Language*. SpringerBriefs, 2018.

Abstract: *Popular deep neural networks for languages processing comprehensively reviewed and explained from a practical perspective. Language modelling, vector representations, and challenging tasks such as WSD, and sentence embeddings investigated to illustrate the used of these networks. This book is packed with valuable advice and experiences obtained from practical development and implementations in the up-and-coming Julia programming language for technical computing. It is a solid introduction to one of the most exciting new areas of natural language processing and computational linguistics.*

(Conference Paper) NovelPersective . First Draft Written. May tweak model slightly again. Some revision to the software required. To be submitted to ACL 2018 Software and Demonstration Track. Will be part of Part C (advantages of one or the other) of thesis. Will also be presented to users at the Western Australian Regional Science Fiction Convention (Swancon), as part of there non-archival academic stream. Bibliographic details
Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. "NovelPerspective". In: *Pending submission* (2018).

Abstract: *We present a proof of concept tool to allow consumers to subset ebooks, based on the main character of the section. Many novels have multiple main characters each with their own storyline running in parallel. A well known example is George R. R. Martin's "Game of Thrones" novel, and others from that series. The NovelPerspective tool detects which character the section is about, and allows the user to generate a new ebook with only those sections. This gives consumers new options in how they consume their media, allowing them to pursue the storylines sequentially, or skip chapters about characters they find boring. We present two simple baselines, and several machine learning based methods for the detection of the main character.*

(Short Journal Paper) DataDeps.jl 2nd Draft Written. To be submitted to JMLR Software Track will be part of Part D (tooling), or appendix of thesis. This will also be the basic of a presentation "Foundational Tools for Data Driven Research, with Applications in NLP and ML" to be submitted to the (non-archival) Julia Language Conference (JuliaCon). Bibliographic details
Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. "DataDeps.jl: Repeatable Data Setup for Reproducible Data Science". In: *Pending submission* (2018).

Abstract: *We present a framework DataDeps.jl for the reproducible handling of static datasets to enhance the repeatability of software scripts used in the data and computational sciences. DataDeps.jl is a library for the Julia programming language. It is used to automate the data setup part of running software accompanies a paper to replicate a result. This step is commonly done manually, which expends time and allows for confusion. This functionality is also useful for other packages which require data to function. It simplifies extending research software via traditional means of a software dependency, as the extension does not have to worry about ensuring the data is setup for its dependency. DataDeps.jl makes it easier to rerun another authors code, thus enhancing the reproducibility of data science research.*

(Unknown) Nonnegatively weighted neural networks I had begun to prepare a theory paper on this, however I discovered it was very similar to an existing work from serveral years early. While there remain some novelties, it may not be worth continuing at this stage as it will not be a nice fit for the thesis topic. Though it does allow for a nice extension to the the ColorDist paper, it would be nearly completely ML focused rather than NLP focused.

1.2 Completion Plan

The table below shows how the publications from Section 1.1, and the key dates from Section 1.2.1, together make a plan to result in a completed dissertation. Note that it is non-chronological, see the chronological time-line in Section 1.2.1.

Part	Publication	Publication Status	Finalisation Date
Introduction / Conclusion etc.	—	—	30/7/18
Part A: Literature Review	Chapters from NRoNL Book	Submit 9/3/18	17/6/18
Part B: Bridging Embeddings to Classical	SOWE2BOW	Pub. 2016	
	SOWE2Sent	Pub. 2016	
	Word-Sense Alignment	Pub. 2018	14/5/18
Part C: Embeddings vs Classical	Sentences Meaning Capture	Pub. 2015	
	ColorDist Estimation	In Review	
	NovelPerspective	Submit 23/3/18	28/5/18
Part D/Appendix	DataDeps.jl	Submit 15/4/18	24/8/18
Overall	—	—	**8/9/18**

The majority of the publications for the thesis are now written. Some still need some editing before they are submitted. They will be grouped into the 4 parts (A-D), which will have introductions and conclusions written for them. The finalisations of these parts will be the predominant task for the remainder of the candidature. The dates given for finalisation here, and in the detailed time-line include the time taken to receive and incorporate supervisor feedback.

1.2.1 Detailed Time-line

Many of the events shown in this timeline are for my own planning purposes. Beyond the events marked as tasks (★ and ✪), The remainder are for informational purposes, and my own planning, e.g. as they impact upon my availability for the points that are core to the completion plan.

Key For the symbols in the Timeline

★ Task with externally mandated deadline

✪ Task with self-imposed deadline

✈ Conference or similar restriction on availability.

Timeline for the remainder of candidature.

	Task/Event	Deadline
★	Neural Representations of Natural Language Book, submitted to SpringerBriefs.	9/3/18
★	NovelPersective Conference Paper submitted to ACL 2018 Demo. Track.	23/3/18
✈	CICLing Conference	18-24/3/18
✈	Swancon Convention/Easter Break	29-2/3/18
✪	DataDeps.jl submitted to JMLR Software Track	15/4/18
✪	Dissertation Document Created, with papers all in place, but blank introductions.	29/4/18
★	JuliaCon Conference abstract due	30/4/18
✪	Dissertation Part B (bridging representations) preparation/finalisation complete.	14/5/18
✪	Dissertation Part C (advantages of one or the other) preparation/finalisation complete.	28/5/18
✪	Dissertation Part A (literature review) written based on book	17/6/18
✈	ACL Conference (if accepted)	15-19/7/18
✪	Dissertation Overall Introduction/Conclusion written	30/7/18
✈	JuliaCon Conference (if accepted)	7-11/8/18
★	Nomination of Examiners submitted to GRS	18/8/18
✪	Dissertation Part D (tooling)/Appendix preparation/finalisation complete.	24/8/18
★	Thesis submitted	8/9/18

Signed ..... (student) ..... (supervisor)

1.3 Problems Encountered So Far

- Supervisor Dr Wei Liu left on maternity leave at the end of 2016 until start of 2018. This meant complete loss of her important input on research for 3 months, and only limited ability to get her input for the remainder of the year. This resulted in one project needing to be abandoned after 1 month of work as it required that input; and more generally, Dr Liu's feedback is a very valuable part of the research training program for my studies.

- The NeCTAR allocation which I had been using as the primary compute for my research was due to expire on the 13th of September 2017. However, due to the Pawsey centre's unexpected de-federation, it was instead terminated on the 5th of July 2017, which severely interrupted the middle of a larger series of work. (I was further expecting that I would be able to renew it, as I had previously for 12 more months after September) It was not possible to move to the new Nimbus cloud (should I have succeeded in an application to that) as Nimbus does not support the Swift Object Storage which I was making heavy use of. The computations had to be closed off, and I still do not have access to compatible Swift Object Storage. As such, when it becomes necessary to to extend upon that work, time will have to be spent rewriting a nontrivial portion of the research software.

- A repeated series of interruptions to computation has occurred due to the multiple power failures in the EE building in the last 12 Months. These have been due to the storm in August, the constructions works from October onwards, and more recently unexplained power glitches. Due to failures of communication, even the "planned" power outages were unexpected and resulted in the termination of long running computations.

- The Neural Representations of Natural Language Book was not a planned publication. It was an unexpected opportunity to create a book (on the request of the publisher), at what is an excellent time for such a book to come into existence; while also creating the literature review required for the thesis. The more extensive time required for the more extensive work, however, did delay other works.