PhD Research Proposal

Lyndon White

14th September 2015

School of Electrical, Electronic and Computer Engineering
Faculty of Engineering, Computing and Mathematics
The University of Western Australia

A. Project Title and Summary

A.1. Project Title

Semantic Vector Representations of Natural Languages

A.2. Summary

The research will investigate methods for the production and utilization of vector representations of natural language preserving meaning. Algorithms producing vector embeddings of sentences and longer documents currently exist, however the field is still developing. Existing methods have not been shown to sufficiently preserve meaning in the vector representation. There has also only been limited investigation in to reversing the projection and resynthesis text from the embedding space. The aims of this project are thus:

- Develop a method for producing semantically consistent vector representations of sentences. This includes evaluating and extending existing methods, and developing new ones.
- Develop a method for resynthesizing text from such vector embeddings. This may be in the form of an extension of a current methods, if they meet the previous aim, or developing new algorithms with the capacity inherent.
- Utilize algorithms in the vector space, to carry out tasks in the natural language space.

B. Research Project

B.1. Background

Over the last five years, word-embeddings have revolutionized Natural Language Processing (NLP). A word embedding is the conversion of a word, into a vector in semantic space. This has several applications and is used to achieve the state of the art solutions to many NLP problems. More recently sentence embeddings have attracted some attention. Sentence Embeddings have also produced state of the art results in there application area. This project aims to create semantically consistent sentence embeddings suitable for using in Natural Language Understanding and Generation (NLU and NLG).

NLP is a key area for modern development. Vast amount of information exists written, or spoken, in natural languages such as English and Chinese, NLP is concerned with processing this. As the amount of information constantly grows, so to does the need to be able to process it computationally. By embedding sentences into a vector space, spacial methods and intuitions can be applied to this processing problem. NLU is the subfield of NLP concerned with creating software which can (to some extend) comprehend the meaning of natural language input. NLG is the field concerned with using software to produce natural language output. Embedding and resynthesizing sentences into and from the vector spaces can be applied to NLU and NLG problems respectively. This combination adds a new angle of attack upon an array of current problems.

Full cycle vector embeddings of sentences would be able to accomplish many tasks which currently require manual intervention. An example how they can be used for abstractive summarization is shown in Figure B.1. Other tasks which could be performed similarly include: paraphrase generation, machine translation and creating descriptions from images. However, currently only limited progress has been made towards the resynthesis step. With just the embedding step state of the art results have been reached in the correspondent tasks: Extractive summarization[1, 2], paraphrase detection [3], similarity measurement for machine translation purposes[**?**], and identifying images based on description[4].

B.2. Problem Statement

Word Embeddings are reversible – it is possible to convert back from a embedding vector to the most similar word. This reversibility is essential for many applications. Often the reversibility is acheived via a nearest neighbor search via embedding the whole vocabulary. The current methods for phrase embeddings are not reversible. Arbitrary phrase vectors can not trivially be converted into a natural language sentence. A key requirement for being able to synthesize a sentence from a sentences vector is for those vectors to be semantically consistent in the first place. Recent results have indicated that current methods may not be sufficiently consistent in their mapping from meaning to position to meet these requirements. This proposal is to devises new methods which do, and to use them to for bidirectional conversion between vectors and sentences.

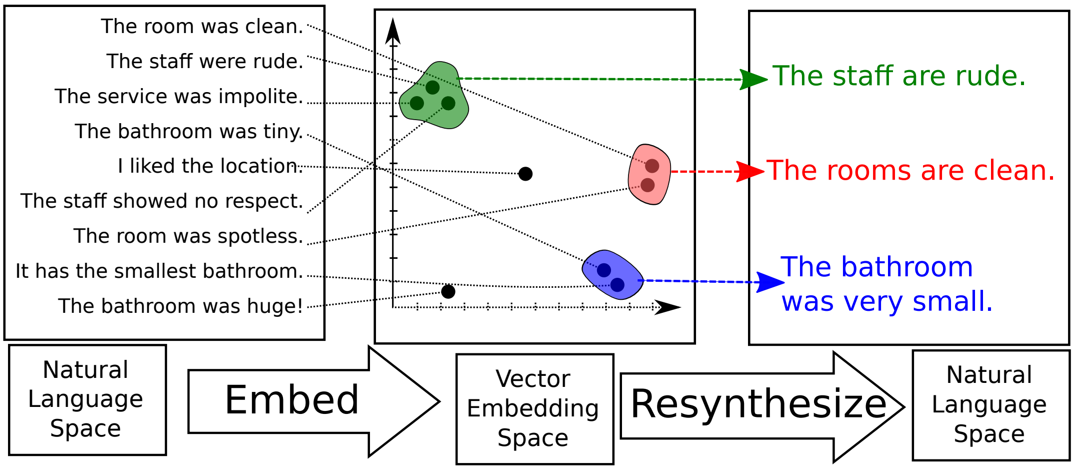


Figure B.1: Work flow for how embeddings may be used to perform abstractive summarization. Sentences (in this case hotel reviews), are taken into the vector space (shown in 2 dimensions for readability, in actuality 50–300 dimentions) where spacial methods are used to cluster commonly occurring meanings, and to disregard outliers. In the Resynthesize step, new sentences are generated which surmise the meaning of each cluster.

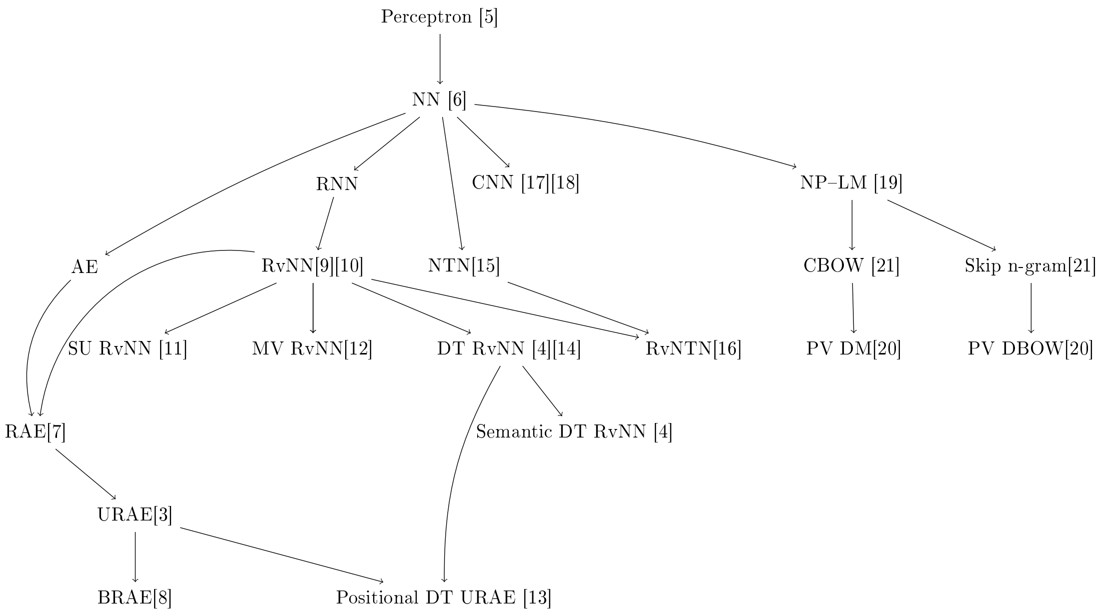


Figure C.1: The "Family Tree" NLP neural networks. The abbreviations are exanded upon in the text below (see **bold** markings)

B.3. Research Question

How can the meaning of a sentence be represented as a vector, such that a vector can be resynthesised into a synonymous sentence.

B.4. Significance

The production of such reversible embeddings is will enhance current NLP techniques to allow for whole phrases to be handled as vectors. Further as other methods for solving word embedding problems – such as short phrase embeddings, and word-sense embeddings – are developed, the extention the reversible phrase embedding methods proposed here will to be obvious and beneficial – as the proposed methods build upon the existing word embedding technologies.

C. Literature Review

Figure C.1 shows a rough outline of the development of the array of methods used for generating embeddings for natural langauge processing. In the following sections the methods are broken down by application.

Year	Author	Method	Data	
2011	Socher et. al.[10]	RvNN	Wall Street Journal sections of Penn Treebank	
2013	Socher et. al.[11]	CVG on RvNN		
2013	Socher et. al.[11]	CVG on SU-RvNN		

Table 1: The application of RvNN decended technologies to parsing. In all cases the Test and Training data was the Wall Street Journal Sections of the Penn Treebank.

Year	Author	Method	Data	Perfo manc
2011	Socher et. al.[7]	Semi-supervised RAE	Movie Reviews[22]	Acc: 77
			Opinions[1]	Acc: 86
2013	Socher et. al. [?]	Recursive NTN	Stanford Sentiment Treebank	Acc: 87
2014	Kalchbrenner et. al.[23]	Dynamic CNN		Acc: 86
2014	Le and Mikolov[20]	PV-DBOW + PV-DM		Acc: 87
2014			IMDB Dataset	Acc 92.58
2015	Zhang and LeCun[24]	Temporal CNN on characters	Amazon Reviews	Acc 95.07

Table 2: The application of various model to the Polarity Sentiment Analysis task. For this task a correct results is limited to determining wether the statement is negitive or positive.

C.0.1. Parsing: RvNN

While parsing is a syntactic task, rather than a semantic one this proposal is concerned with, it was the first task to which the Recursive Network (**RvNN**) was applied to. Table 1 shows the performance of the methods. The RvNN generalizes the reused of the output as an input, which is present in the Recurrent Neural Network (**RNN**) to be performed over a tree of inputs, with each layer merging into the next.

C.0.2. Sentiment Analysis: RAE, URAE, PV-DM, PV-DBOW

C.1. Scholars in the Field

- A/Prof. Phil Blunsom, Department of Computer Science, Univerity of Oxford. UK. Email: phil.blunsom@cs.ox.ac.uk
- Dr Fei Liu, School of Computer Science, Carnegie Mellon University, USA. Email: feiliu@cs@cmu.edu[2]

[2] Dr Fei Liu will be moving to the University of Central Florida in the very near future. Her future contact details are thus expected to change.

Year	Author	Method	Evaluation Task	
2013	Sorcher et. al. [?]	Recursive NTN	Sentiment Analysis: Exact Score	Stanfor
2014	Le and Mikolov[20]	PV-DBOW + PV-DM	Sentiment Analysis: Exact Score	Stanfor
2014	Kalchbrenner et. al.[23]	Dynamic CNN	Sentiment Analysis: Exact Score	Stanfor
2015	Zhang and LeCun[24]	Temporal CNN on characters	Sentiment Analysis: Exact Score	A
2012	Socher et. al.[12]	MV-RvNN	Sentiment Analysis: Exact Scrore	M

Table 3: The application of various model to the Polarity Sentiment Analysis task. For this task a correct results is limited to determining wether the statement is negitive or positive.

- Dr Richard Socher, Computer Science Department,Stanford University, USA. Email: richard@socher.org
- Dr Tomas Mikolov, Facebook AI Research, USA. Email: tmikolov@fb.com
- Prof. Yann LeCun, Computer Science Department, New Your University, USA. Email: yann@cs.nyu.edu
- Prof. Yoshua Bengio, Department of Computer Science and Operations Research, Canada, Email: yoshua.bengio@umontreal.ca

References

[1] M. Kågebäck, O. Mogren, N. Tahmasebi, and D. Dubhashi, "Extractive summarization using continuous vector space models," in *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)@ EACL*, 2014, pp. 31–39. [Online]. Available: http://www.aclweb.org/anthology/W14-1504

[2] D. Yogatama, F. Liu, and N. A. Smith, "Extractive summarization by maximizing semantic volume," *Conference on Empirical Methods in Natural Language Processing*, 2015. [Online]. Available: http://www.cs.cmu.edu/~dyogatam/papers/yogatama+liu+smith.emnlp2015.pdf

[3] R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, and C. D. Manning, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection," in *Advances in Neural Information Processing Systems 24*, 2011. [Online]. Available: http://www.socher.org/uploads/Main/SocherHuangPenningtonNgManning_NIPS2011.pdf

[4] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014. [Online]. Available: https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/download/325/45

[5] F. Rosenblatt, "Principles of neurodynamics. perceptrons and the theory of brain mechanisms," DTIC Document, Tech. Rep., 1961.

[6] D. E. Rumelhart, G. E. Hintont, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, p. 9, 1986. [Online]. Available: http://www.cs.toronto.edu/~hinton/absps/naturebp.pdf

[7] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-supervised recursive autoencoders for predicting sentiment distributions," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011. [Online]. Available: http://www.socher.org/uploads/Main/SocherPenningtonHuangNgManning_EMNLP2011.pdf

[8] J. Zhang, S. Liu, M. Li, M. Zhou, and C. Zong, "Bilingually-constrained phrase embeddings for machine translation." ACL, June 2014. [Online]. Available: http://anthology.aclweb.org/P/P14/P14-1011.pdf

[9] J. B. Pollack, "Recursive distributed representations," *Artificial Intelligence*, vol. 46, no. 1â"2, pp. 77 – 105, 1990. [Online]. Available: http://www.sciencedirect.com/science/article/pii/000437029090005K

[10] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng, "Parsing natural scenes and natural language with recursive neural networks," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 129–136. [Online]. Available: http://nlp.stanford.edu/pubs/SocherLinNgManning_ICML2011.pdf

[11] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with compositional vector grammars," in *ACL*, 2013. [Online]. Available: http://www.socher.org/uploads/Main/SocherBauerManningNg_ACL2013.pdf

[12] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic compositionality through recursive matrix-vector spaces," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* Association for Computational Linguistics, 2012, pp. 1201–1211. [Online]. Available: http://www.aclweb.org/anthology/D12-1110

[13] M. Iyyer, J. Boyd-Graber, and H. D. III, "Generating sentences from semantic vector space representations," in *NIPS Workshop on Learning Semantics*, 2014. [Online]. Available: http://cs.umd.edu/~miyyer/pubs/2014_nips_generation.pdf

[14] M. Iyyer, J. Boyd-Graber, L. Claudino, R. Socher, and H. Daumé III, "A neural network for factoid question answering over paragraphs," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 633–644. [Online]. Available: https://cs.umd.edu/~miyyer/pubs/2014_qb_rnn.pdf

[15] R. Socher, D. Chen, C. D. Manning, and A. Y. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Advances in Neural Information Processing Systems 26*, 2013. [Online]. Available: http://nlp.stanford.edu/~socherr/SocherChenManningNg_NIPS2013.pdf

[16] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, vol. 1631. Citeseer, 2013, p. 1642. [Online]. Available: nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf

[17] L. E. Atlas, T. Homma, and R. J. Marks II, "An artificial neural network for spatio-temporal bipolar patterns: Application to phoneme classification," in *Proc. Neural Information Processing Systems (NIPS)*, 1988, p. 31. [Online]. Available: http://goo.gl/4OApDV

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. [Online]. Available: http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf

[19] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, "Neural probabilistic language models," in *Innovations in Machine Learning.* Springer, 2006, pp. 137–186. [Online]. Available: http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

[20] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188–1196. [Online]. Available: http://jmlr.org/proceedings/papers/v32/le14.pdf

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013. [Online]. Available: http://arxiv.org/pdf/1301.3781v3

[22] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 2005, pp. 115–124. [Online]. Available: http://ssli.ee.washington.edu/conferences/ACL2005/ACL/pdf/ACL15.pdf

[23] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014. [Online]. Available: http://nal.co/papers/Kalchbrenner_DCNN_ACL14

[24] X. Zhang and Y. LeCun, "Text understanding from scratch," *CoRR*, vol. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2015. [Online]. Available: http://arxiv.org/abs/1502.01710

D. Research Project Details

- This project does not involve the collection of confidential or sensitive information.
- The are no current intellectual property agreements relating to this research. There are no plans to commercialize the products of this research during the duration of the candidature.
- This project does not involve fieldwork.

- Additional computational resources are required and have been acquired for the project. For purposes of experimentation and development of the algorithms distributed computing methods are used. The estimated requirements are 32 CPU-Cores and at least 2Gb RAM per core. Further more. the computers need to be collocated on a highspeed network. A successful application has been made to NeCTAR for this allocation. The allocation will need to be renewed on the 31st of December 2015.

- There are no advanced statistical analysis required for this project, beyond the consideration and design of the algorithms developed as statistical analysis tools themselves.

- The outputs of the project will be communicated by publication in journals and conferences. As is conventional within this area, reseach will be primery communicated though conference papers. The intent is to publish one conference paper in the first year, two in the second, and 1 in the third. As judged appropriate based on research in question, one or more of these conference papers will be extended into a journal article. The thesis shall be presented as this series of conference and journal papers, with additional introductory and concluding chapters.

- This project does not require any approvals. It has been investigated if the intended research publications require Defense Export Control Office (DECO) approval for publications. DECO approval is not required for publication.

- Any new and derived data sets will be placed in the Signals and Information Processing lab's existing UWA Institutional Research Data Store (IRDS). Where licensing allows it will also be publicly published though the candidates website. The focus of this course of study is to create new methods of processing data rather than new data. The methods for processing data will be version controlled via private Github, with intent to open source them at the pertinent times.

In the table below is shown a skills audit for the skills required in this project

Professional and Research Skills	Rating				Evidence	Plan for Acquisition
	None	Basic	Competent	Proficient		
Understanding and application of data collection and analysis methods			C		Completion of Honours project, which involved collection of large amounts of data, and its analysis.	Not Required

Identifying and accessing appropriate bibliographic resources			C		Annotated bibliography maintained. Completed Honours project. Completed UWA Library “Keeping Up to Date” workshop	Not Required
Understanding of mathematics required for this area (Probability, Linear Algebra)			C		Completed Pure Mathematics Major, as part of BCM	Not Required.
Use of programming languages for this area (Matlab, Python, Julia)				P	Completed Computation Major, as part of BCM Experience as professional software developer	Not Required
Use of signal processing techniques			C		Completed Electrical and Electronic Program as part of BE	Not Required
Use of Distributed Computing Resources (MPI etc)			C		Completed Developer Training at Pawsey Super Computer Center.	Complete
Principles and conventions of academic writing		B			Completion of Honours. However, this took intensive editing.	Attend GRS and Library Writing Workshops
Self discipline and motivation			C		Have worked at lower paying, much less enjoyable jobs to get to university.	Not Required
Time and project management			C		Completion of Honours. Completion of heavily project assessed Computer Science and Engineering Majors Including 4 project management units.	Not required
Awareness of issues relating to intellectual rights			C		Attended Graduate Research School Induction Session on Scholarly Ethics. Read the UWA Code of Ethics.	Not Required
Ability to constructively defend research outcomes at presentations			C		Have presented my Honours at school symposium. Have presented school seminar.	Not Required

E. Timeline: Research Training and Academic Tasks

The timeline for this research program is spread over 3 years, to inline with the candidate's Australia Post Graduate Award (APA) duration. If particular difficulties arise, the APA can have a 6 month extension, this also is indicated in the timeline below, to allow for adjustment to be scaled. Failure to complete before the termination of the funding will result in sever difficulties to the candidate's living circumstance and will likely result in non-completion.

This timeline should be reach in conjunction with the Candidature Tasks, on the cover-sheet, and the Research Program Overview in the at the end of this document. Key milestones are marked in *italics*.

01/03/2009 Academic Conduct Essentials (candidature task)

08/03/2015 *Enrollment*

24/04/2015 GRS: Theses and Publications Seminar

04/06/2015 GRS: Research Skills Workshop

05/06/2015 Pawsey Supercomputer Training (candidature task)

22/06/2015 GRS: How to Write a Research Proposal (at weekend writing retreat)

16/07/2015 Library Workshop: Keeping Up to Date [with Literature]

18/09/2015 *Research proposal*

08/03/2016 *Annual report Year 1*

08/03/2016 *Confirmation of candidature*

15/02/2017 *Nomination of Examiners*

08/03/2017 *Annual report Year 2*

08/02/2018 Dissertation Draft submitted to supervisors

08/03/2018 *Dissertation submitted for examination*

08/03/2018 APA Termination

05/09/2018 APA Extension Termination

Description	Costs			Source	
	Year 1	Year 2	Year 3	School	GRS
Administrative and Research Costs					
Workstation	$1500	$0	$0	$1500	$0
Linguistic Data Consortium Membership	$0	$2400	$0	$2400	$0
Training Costs					
GRS/Library Seminars and Workshops	$0	$0	$0	$0	$0
Statistics Training Course	$0	$198	$0	$198	$0
Conference Attendance					
Domestic: Flights, Registration, Accommodation	$1500	$0	$1500	$3000	$0
International: Flights, Registration, Accommodation	$0	$2000	$0	$150	$1850
Subtotal:	$3000	$4598	$1500	$7248	$1850
				Total:	$9098

Table 5: Budget

F. Budget

The budget for this research program is detailed in Table 5. The most significant cost of the project is the purchasing of a 2016 membership to the Linguistic Data Consortium. This membership allows the obtaining of the vast majority of the LDC data-sets at no additional cost. The piece-wise cost for the key data sets required for this research, Gigaword v5 and Penn Treebank-3, would otherwise cost $6,000 and $1,500 respectively. As this is an institution wide membership, it will also allow the group to obtain and update many other data-sets used for other projects. It was determined to obtain membership in the second year of the project, rather than the first to ensure best utilization.

G. Supervision

Principal & Coordinating Supervisor: Professor Roberto Togneri (40%)

- Provide expertise in spoken language systems, statistical signal processing and pattern recognition.
- Directing overall research training program
- Reviewing research outputs
- Provide regular feedback, on both overall, and current subproject progress

Co-Supervisor: Dr Wei Liu (40%)

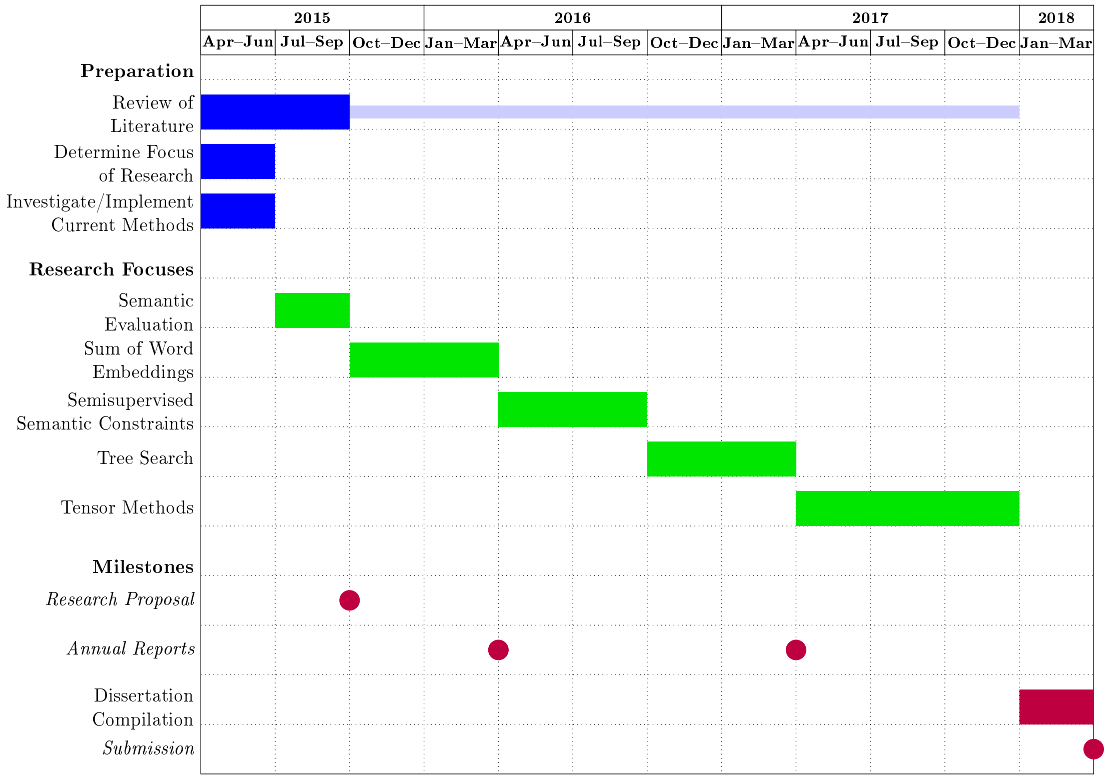
- Provide expertise in natural language processing, and the conventions of publication in the field.

- Reviewing research outputs
- Provide regular feedback, on both overall, and current subproject progress

Co-Supervisor: Winthrop Professor Mohammed Bennamoun (20%)

- Provide expertise in machine learning, particularly in deep neural systems
- Reviewing research outputs
- Provide regular feedback, on both overall, and current subproject progress

H. Research Program Overview


Apr–Jun
Jul–Sep
Oct–Dec
Jan–Mar
Apr–Jun
Jul–Sep
Oct–Dec
Jan–Mar
Apr–Jun
Jul–Sep
Oct–Dec
Jan–Mar
Preparation
Review of Literature
Determine Focus of Research
Investigate/Implement Current Methods
Research Focuses
Semantic Evaluation
Sum of Word Embeddings
Semisupervised Semantic Constraints
Tree Search
Tensor Methods
Milestones
Research Proposal
Annual Reports
Dissertation Compilation
Submission