Revisions - Assessment of Dense Word Representations for

Text Classification in Biocuration of Infectious Disease

Daniel E. Sullivan

# ABSTRACT (Academic)

Text mining researchers have applied a range of techniques to analyze the biomedical literature. This research addresses the problem, can unsupervised learning generate a representation that improves on the commonly used term frequency-inverse document frequency (TF-IDF ) representation by capturing semantic relations? Semantic vectors are a more recently developed representation that demonstrates high levels of performance in classification and analogical reasoning tasks. The original motivation for this research was the need to support biocuration efforts in the PATRIC project, but the findings of this research generalize to apply to text classification, clustering, and analogical reasoning. The analysis measures the quality of sentence classification using term TF-IDF representations, and finds a practical upper limit to precision and recall in a biomedical text classification task (F1-score of 0.85). TF-IDF representations do not capture semantic relations between terms, and this may contribute to limits on classification performance. Arguably, one could use ontologies to supplement TF-IDF, but ontologies are sparse in coverage and costly to create. This prompts a correlated question: can unsupervised learning capture semantic relations at least as well as existing ontologies, and thus supplement existing sparse ontologies? Semantic vectors are designed to capture the semantics of words by using vectors which are proximity to the vectors of semantically related words. A shallow neural network implementing the Skip-Gram algorithm is used to generate semantic vectors using a corpus of approximately 2.4 billion words. The ability to capture meaning is assessed by comparing semantic vectors generated with MESH. Results indicate that semantic vectors trained by unsupervised methods capture comparable levels of semantic features in some cases, such as amino acid (92% of similarity represented in MESH), but perform substantially poorer in more expansive topics, such as pathogenic bacteria (37.8% similarity represented in MESH). Possible explanations for this difference in performance are proposed along with a method to combine manually curated ontologies with semantic vector spaces to produce a more comprehensive representation than either alone. Semantic vectors are also used as representations for paragraphs, which, when used for classification, achieve an F1-score of 0.92. The results of classification and analogical reasoning tasks are promising but a formal model of semantic vectors, subject to the constraints of known linguistic phenomenon, is needed. This research includes initial steps for developing a formal model of semantic vectors based on a combination of linear algebra and fuzzy set theory subject to the semantic molecularism linguistic model. This research is novel in its analysis of semantic vectors applied to the biomedical domain, analysis of different performance characteristics in biomedical analogical reasoning tasks, comparison semantic relations captured by between vectors and MESH, and the initial development of a formal model of semantic vectors.

# Narrow the focus of the Introduction

Chapter 1

The focus of the research described here is to evaluate the use of dense word representations derived by unsupervised learning algorithms. The evaluation will consist of measuring the quality of sentence classification using term frequency-inverse document frequency (TF-IDF) representations, a statistical approach that does not take into account the semantics of word or phrases. [Salton, 1975] An alternative representation scheme using dense vectors, referred to as semantic vectors in this research, is evaluated. Semantic vectors are designed to capture some semantics of words by using vectors which are in close proximity to the vectors of semantically related words. The ability to capture meaning is assessed by comparing semantic vectors generated using unsupervised learning techniques with manually curated ontologies. Semantic vectors are also used to generate vector representations for paragraphs. These vector representations are evaluated using a classification task. Furthermore, this research includes initial steps for developing a formal model of semantic vectors based on a combination of linear algebra and fuzzy set theory. This research is novel in its objective of focusing on biocuration for infectious bacterial diseases and evaluating dense word representations for each task and comparing results to commonly used existing techniques. This proposed research has three specific aims.

**Aims of Research**

There are three primary aims of this research:

* Evaluate text classifiers using TF-IDF representation and multiple machine learning algorithms.
* Evaluate semantic vector representations and their ability to capture word meaning, which is measured by comparing automatically generated word vectors with comparable terms in manually curated ontologies.
* Evaluate the quality of paragraph classifiers using semantic vectors for paragraphs.

In addition the three specific aims, this research begins to outline a formal model of semantic vectors using linear algebra and fuzzy set theory. This is a novel approach to semantic vectors; the author is unaware of any similar attempts to combine these two formalisms and apply them to linguistic phenomenon.

This research also provides an analysis of characteristics of dense word vectors and clusters of dense word vectors and their relation to the quality of classification and information extraction. Together, these specific aims will enhance the understanding of dense word vectors as a representation scheme for improving the quality of classification and information extraction.

The motivation for this research is the hypothesis for the need of broad, semantic representations of biomedical terms, which has not yet been met. As noted by one researcher:

Broad coverage semantic taxonomies such as WordNet (Felbaum, 1998) and CYC (Lenat, 1995) have been constructed by hand at great cost; while a crucial source of knowledge about the relation between words, these taxonomies still suffer from sparse coverage.[Snow, 2006]

Many text mining applications use the term frequency inverse document frequency (TF-IDF) document model. While useful in many areas, the conventional implementation of TF-IDF based classifiers fail to take into account semantic information contained with a text. The overall objective of this research is to assess the value of representations that include semantic features when applied to word and paragraph representations.

# 

# Why was a Shallow Neural Network chosen over a Deep Neural Network

Chapter 3

This research uses a shallow neural network that implements the Skip-Gram algorithm to learn semantic vectors for words and paragraphs. Recent advances in machine learning have repeatedly demonstrated that deep, multi-layer networks can achieve high, and sometimes state of the art, levels of performance. [Le, 2013], [Glorot, 2011], [Collobert, 2008]. There are, however, drawbacks to using multi-layer networks and the decision to use a two layer network was based on a consideration of the benefits and costs of two layer vs multi-layer models.

The shallow Skip-Gram model has several advantages.

First, Skip-Gram, as implemented in the word2vec program [Mikolov, 2013a], has been widely successfully used in the text processing community for tasks including: sentiment analysis [Zhang, 2015], bilingual semantic representations [Wolf, 2014], and extracting quantitative variables from unstructured text [Amunategui, 2015]. The experiments described in this dissertation demonstrate high levels of performance in text classification tasks (F1-score > 0.9). In addition, the original Skip-Gram algorithm has been improved by introducing negative sampling, increasing the performance of the algorithms without increasing the time to train the network. [Mikolov, 2013b] The high levels of performance in a range of text analysis areas support the hypothesis that Skip-Gram is robust and applicable to a range of text analysis problems.

Deep neural networks have been used to train semantic vectors ([Collobert, 2004], [Turian, 20100], [Huang, 2013]) but they are computationally more expensive than the Skip-Gram algorithm. [Mikolov, 2013a]. Computational complexity is an important consideration given the size of trainings sets, which in the case of this study, included training the network on a corpus of over 2 billion words. There is the potential for a deep network to improve the performance of a classifier based on performance of other deep learning networks. However, one would also have to investigate if similar or better performance increases could not be achieved by increasing the size of the training corpus or number of training epochs while still training the shallow network in less time than required to train a deep network.

Deep learning networks do not always find optimal solutions and sometimes yield sub-optimal results. [Pandey, 2014]. Heuristics are used to tune a number of configuration parameters, such as the number of training epochs, learning rate, momentum, and batch size but some researchers have found that poor choices for these hyper-parameters can lead to decreased performance. [Bergstra, 2011]. Skip Gram requires hyper-parameter tuning as well but given the lower computational complexity, it is more efficient to apply grid search to find optimal hyper-parameter configurations.

Finally, it has been demonstrated that wide, shallow networks can perform better at some classification tasks than deep networks. [Pandey, 2014] Specifically,

Using a single RBM [Restricted Boltzman Machine] to learn a wide layer, we are able to obtain better results for many classification tasks than obtained by multi-layer neural network initialized using a deep belief network and fine-tuned using backpropagation. [Pandey, 2014]

Pandey’s work does not compare Skip-Gram to a deep network but it does demonstrate that deep neural network are not inherently better at classification than shallow networks. More research is needed in this area to identify what kinds of classification problems are well suited to shallow but wide neural networks, such as used by Skip-Gram, and which are better suited for deep neural networks.

**References**

Amunategui, Manuel, Tristan Markwell, and Yelena Rozenfeld. "Prediction Using Note Text: Synthetic Feature Creation with word2vec." *arXiv preprint arXiv:1503.05123* (2015).

Bergstra, James S., et al. "Algorithms for hyper-parameter optimization."*Advances in Neural Information Processing Systems*. 2011.

Collobert, Ronan, and Jason Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning."*Proceedings of the 25th international conference on Machine learning*. ACM, 2008.

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Domain adaptation for large-scale sentiment classification: A deep learning approach." *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011.

Huang, Eric H., et al. "Improving word representations via global context and multiple word prototypes." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012.

Le, Quoc V. "Building high-level features using large scale unsupervised learning." *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013.

Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013a).

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013b.

Mnih, Andriy, and Geoffrey E. Hinton. "A scalable hierarchical distributed language model." *Advances in neural information processing systems*. 2009.

Pandey, Gaurav, and Ambedkar Dukkipati. "To go deep or wide in learning?."*arXiv preprint arXiv:1402.5634* (2014).

Turian, Joseph, Lev Ratinov, and Yoshua Bengio. "Word representations: a simple and general method for semi-supervised learning." *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 2010.

Wolf, Lior, et al. "Joint word2vec networks for bilingual semantic representations." *International Journal of Computational Linguistics and Applications* 5.1 (2014): 27-44

Zhang, Dongwen, et al. "Chinese comments sentiment classification based on word2vec and SVM perf." *Expert Systems with Applications* 42.4 (2015): 1857-1863.

# Why were the particular methods chosen? Describe MESH terms and add more detail about the comparisons.

**Chapter 3**

This research includes a comparative analysis of several methods for representing text and comparing semantic representations, including: using term frequency-inverse document frequency (TF-IDF) and semantic vectors for text representations, using machine learning algorithms, and ontology comparisons specifically developed for this research. The justification of each choice follows below.

TF-IDF representations was described by Salton in 1975 and has since been cited over 6,300 times according to Google Scholar. [Salton, 1975] The representation has been successfully used in text classification tasks, including short text classification [Robertson, 2004], news filtering [Lang, 2002], and spam filtering [Drucker, 1999]. These tasks are sufficiently similar to biomedical classification to warrant the use of TF-IDF. In addition, the combination TF-IDF combined with support vector machines has been studied as well. [Joachims, 1998] [Leopold, 2002].

Semantic vectors can potentially improve on TF-IDF as a feature representation scheme. Specifically, semantic vectors capture semantic relations between words. It is hypothesized that using a representation that captures semantic relations will improve the quality of classification. Results of experiments described in this dissertation support that hypothesis. Semantic vectors have used for sentiment analysis [dos Santos, 2014], part of speech tagging [Tsuboi, 2014], and word sense disambiguation [Chen, 2014]. This breadth of task indicates that semantic vectors are sufficiently robust to support a range of text analysis tasks.

Several machine learning algorithms are used in this study. Algorithms were chosen to represent several widely used classes of learners. Linear learners include support vector machines (SVMs), Percptron, and Linear Regression. SVMs are wide margin classifiers widely used in text classification tasks. [Hearst, 1998]. Classification tree methods used include decision trees and random forests. Naive Bayes was the one probabilistic learning algorith used. The ensemble algorithm, Adaboost, was used as well. These algorithms were selected to help identify the optimal combination of representation (TF-IDF or semantic vector) and classification algorithm. It is not clear, *a priori*, that any one type of classification algorithm would work best with semantic vectors. The fact that SVMs with a linear kernel perform well on TF-IDF classifications support the hypothesis that text representations based on words are linearly separable.

Ontology comparisons are used to evaluate the ability of semantic vectors to capture semantic relations. Previous studies have depended on common knowledge analogies, such as countries and their capitals, to evaluate the ability of semantic vectors to capture useful semantic relations. One of the objectives of this research is to evaluate the quality of semantic relations captured by semantic vectors trained on a large, biomedical corpus, with particular emphasis on infectious disease topics. Biomedical ontologies such as MESH, represent important biomedical concepts and relations as defined by expert curators. [Lipscomb, 2000] Designers of MESH note that the resources complements automatic text analysis methods by providing precision not typically found in automatic information retrieval methods:

Even with advances in automation and resulting changes in the capabilities of indexing and searching, an important role remains for MeSH in organizing information in a way that provides precision and power in retrieval. [Lipscomb, 2000]

For this study, MESH is considered the “gold standard” of semantic representations. MeSH has existed since 1960 and has been described as “one of the most sophisticated thesauri in existence today.” [Nelson, 2001] MeSH terms represent single concepts within the biomedical field. MeSH terms comprise 15 hierarchies, know as the MeSH Tree Structure. [Lowe, 1994] Each hierarchy constitutes a a structured set of increasingly specific terms. In addition to biomedical terms, there are additional terms to support search, such as publication types. These additional terms are not relevant to the research at hand.

MESH is a thesaurus that encompasses features of an ontology, including the ability to reason about the ontology as a graph. Nodes of the graph represent terms and measures such as the number of edges between nodes can be uses as a measure of semantic similarity. Semantic vectors, however, exist in a linear space and similarity is typically measured using cosine or Euclidean distance. Obviously, direct comparisons between the two representations are not possible. Instead, custom methods were developed to compare similarity measures in MESH and semantic vectors.

Computing semantic similarity between semantic vectors is straight forward: the cosine of two vectors is used as a measure of similarity. Computing the semantic similarity between terms in an ontology such as the MeSH hierarchies, requires a different metric. The evaluations described in this dissertation used the Sanchez information content method. [Sanchez, 2011]. Information content (IC) of a term is defined as the amount of information it conveys in a context; it has been widely used as a measure of semantic similarity. Semantic similarity, in turn, is understood as a degree of taxonomical resemblance. [Goldstone, 1994] The Sanchez method for computing IC is applied to MeSH because it is both scalable and produces high quality results, as reported in [Sanchez, 2011]. The Semantic Similarity Toolkit implementation of the Sanchez measure is used in this research. [Harispe, 2014].

**References**

Chen, Xinxiong, Zhiyuan Liu, and Maosong Sun. "A Unified Model for Word Sense Representation and Disambiguation." *EMNLP*. 2014.

dos Santos, Cícero Nogueira, and Maira Gatti. "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts." *COLING*. 2014

Drucker, Harris, Donghui Wu, and Vladimir N. Vapnik. "Support vector machines for spam categorization." *Neural Networks, IEEE Transactions on*10.5 (1999): 1048-1054.

Goldstone, Robert L. "Similarity, interactive activation, and mapping." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20.1 (1994): 3.

Harispe, Sébastien, et al. "The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies." *Bioinformatics* 30.5 (2014): 740-742.

Hearst, Marti A., et al. "Support vector machines." *Intelligent Systems and their Applications, IEEE* 13.4 (1998): 18-28.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features, Springer.

Lang, Ken. "Newsweeder: Learning to filter netnews." *Proceedings of the 12th international conference on machine learning*. 1995.

Leopold, Edda, and Jörg Kindermann. "Text categorization with support vector machines. How to represent texts in input space?." *Machine Learning*46.1-3 (2002): 423-444.

Lipscomb, Carolyn E. "Medical subject headings (MeSH)." *Bulletin of the Medical Library Association* 88.3 (2000): 265.

Lowe, Henry J., and G. Octo Barnett. "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches." *Jama*271.14 (1994): 1103-1108.

Nelson, Stuart J., W. Douglas Johnston, and Betsy L. Humphreys. "Relationships in medical subject headings (MeSH)." *Relationships in the Organization of Knowledge*. Springer Netherlands, 2001. 171-184.

Robertson, S. (2004). "Understanding inverse document frequency: on theoretical arguments for IDF." Journal of Documentation **60**(5): 503-520.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing." *Communications of the ACM* 18.11 (1975): 613-620.

Sánchez, David, Montserrat Batet, and David Isern. "Ontology-based information content computation." *Knowledge-Based Systems* 24.2 (2011): 297-303.

Tsuboi, Yuta. "Neural Networks Leverage Corpus-wide Information for Part-of-speech Tagging." *EMNLP*. 2014.

# Highlight how data, representation, and algorithms affect outcomes.

**Chapter 4**

The results presented here are influenced by a combination of the data used to train classifiers, the representation of that data, and the machine learning algorithms used to create classifiers.

The data set used to train the classification algorithm constrain the vocabulary that can be learned by semantic vector training algorithms. Words that do not appear in the training corpus will not have semantic vectors generated. This effectively limits the usefulness of the semantic vectors to the domain of the training corpus. In this research, a combination of Medline abstracts and full text PubMed Central papers are used to train semantic word vectors. The corpus included over 2 billion words in context. The corpus includes commonly used English words as well as biomedical and biological terms that are not likely to be found in large number in non-biomedical corpora, such as Wikipedia or Google News. Biocuration in the area of infectious disease research was the initial motivation for this research so the training corpus should be sufficient to create semantic vectors that can be used in application designed to support biocuration and related biomedical tasks.

Representations are used to make explicit features of texts that can be used by machine learning algorithms to create classifiers. TF-IDF captures features of words relative to their distribution in a document and across an entire corpus. Semantic vectors capture semantic and syntactic features of words based on the context in which those words are used. The latter representation present the opportunity to use semantic features which are not captured by TF-IDF. Arguably, one could supplement TF-IDF representations with features derived from biomedical ontologies but such ontologies are sparse and difficult to build. [REF] In some cases, representations cannot manifest a sufficient set of features to allow machine learning algorithms to discover a model that effectively classifies input text into a set of categories. As results from the TF-IDF experiment in this study show, additional training instances cannot overcome insufficient feature representation. Some practitioners resort to the practice of feature engineering, which is the process of creating derived or other explicit features on a case-by-case basis. This can be an effective way to develop a high quality classifier but it does not scale as a general solution.

Machine learning algorithms also impact the outcome of the experiments described here. In general, supervised machine learning programs attempt to maximize some objective function, such as maximizing the distance from margins in a data set, or creating a set of decision trees that together minimize the error rate on predictions over a training set. Machine learning algorithms sometimes have parameters that are not learn; these are referred to as hyper-parameters. Hyper-parameters include learning rates, training set sizes, number of training epochs, and momentum (a parameter that influences the ability of algorithms to avoid local minima).

# Discuss how additional new techniques can be used to improve the quality of semantic vectors.

**Chapter 4**

The semantic vectors used in this research were developed using the Skip-Gram algorithm. The quality of semantic vectors may be improved by employing additional techniques.

As outlined in Chapter 6, section “Representing Knowledge from Ontologies Directly into Vector Space”, the knowledge explicitly represented in ontologies can be used to initialize semantic vectors. This is one way to incorporate the declarative knowledge of ontologies into semantic vectors.

Alternatively, semantic vectors can be combined with classifiers that use TF-IDF representations supplemented with information from ontologies, such as the taxonomic lineage of terms defined in biomedical ontologies. This approach would not improve the quality of semantic vectors directly, but it could improve the overall performance of a classification program.

Advances in deep learning may also improve the quality of semantic vectors, although as noted earlier, there are demonstrated cases in which breadth of a single layer can outperform narrower but deeper neural networks. One weakness of the Skip-Gram algorithm is that a single word with multiple meanings are subject to training from multiple contexts. For example, “bank” can appear in contexts about financial institutions or about rivers. Context vectors have been used with TF-IDF and should be adaptable to semantic vector models as well. [Chen, 200] There has been some success in using context to derive syntactic properties that may be useful in improving the performance of semantic vectors. [Erk, 2008] There has also been progress in capturing both global and local context with neural networks as in [Huang, 2012].

Deep learning techniques are perhaps the most promising avenue to improve the quality of representation. Recurrent neural networks are particularly promising because they can compress the history of previously encountered words in the training stream into a low dimensional space and have the potential to form short term memory, which can capture some aspects of context. [Mikolov, 2010] Convolutional neural networks have also performed well in natural language processing tasks, although these models are better known for their exemplary performance in image processing. [Kalchbrenner, 2014] Recursive neural network have used with morphological features as noted in [Loung, 2013]. Recursive neural networks are computationally more complex than Skip-Gram, but this can capture a broader range of features because it operates at the morphological and not word-based level. For example, “hydrate” and “dehydrate” are often used in similar contexts and therefore have close semantic vectors generated by Skip-Gram. The ability to work at a morphological level would allow a classifier to distinguish the antonym morpheme “de” in “dehydrate” and adjust the semantic vector of both forms of “hydrate.”

In all cases, additional research is need to evaluate the trade-offs between the quality of representation and the computational complexity of the learning algorithm. Also, the literature is inconsistent in the use of some deep learning techniques. For example, [Kim, 2014] demonstrated significantly better results with convolution networks for a sentence classification task than [Kalchbrenner, 2014]. Additional formal analysis of deep learning techniques could help shed light on how to configure well established patterns, such as convolutional and recursive networks. Until then, one should expect extended periods of experimentation to find optimal numbers and types of layers in a deep neural network to solve a particular problem.

**References**

Chen, Keh-Jiann, and Jia-Ming You. "A study on word similarity using context vector models." *Computational Linguistics and Chinese Language Processing* 7.2 (2002): 37-58.

Erk, Katrin, and Sebastian Padó. "A structured vector space model for word meaning in context." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008.

Huang, Eric H., et al. "Improving word representations via global context and multiple word prototypes." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012.

Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. "A convolutional neural network for modelling sentences." *arXiv preprint arXiv:1404.2188*(2014).

Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).

Luong, Thang, Richard Socher, and Christopher D. Manning. "Better Word Representations with Recursive Neural Networks for Morphology." *CoNLL*. 2013.

Mikolov, Tomas, et al. "Recurrent neural network based language model."*INTERSPEECH*. Vol. 2. 2010.

# Discuss how the results of the experiment generalize to a range of tasks

**Chapter 6**

Semantic vectors are a fundamental representation scheme that captures, to some degree, the meaning of words, sentences, and paragraphs. The experiments in this dissertation provide evidence for the hypothesis that a semantically rich representation can improve upon the performance of a text representation scheme that focuses on non-semantic features, such as word frequency. It is reasonable to propose that, as a fundamental representation scheme, semantic vectors should generalize to other tasks. One researcher has clearly argued for the importance of semantic vectors to deep learning applied to natural language processing: “Our results add to the well-established evidence that unsupervised pre-training of word vectors is an important ingredient in deep learning for NLP.” [Kim, 2014]

In addition to text classification, semantic vectors can be applied to other text mining tasks, including part of speech tagging, named entity recognition, and sentence parsing. Researchers were recently able to obtain state of the art performance in part of speech tagging using neural networks and semantic vector representations. [Tsuboi, 2014] Named entity recognition is an especially difficult task in the biomedical domain. [Ananiadou, 2011]. A recent application of semantic vectors to named entity recognition (NER) demonstrated promising results but also found that performance increased with the size of the training set but only up to a certain point. [Siencnik, 2015]. More experimentation is required to understand impact of training set size on NER performance. A recent paper on sentence parsing using convolutional networks, semantic vectors, and variable size convolutional filters demonstrate favorable results with multi-class sentiment prediction. [Yin, 2016]

Just as additional techniques can help improve the quality of semantic vectors, semantic vectors can improve the quality of performance on a number of text mining related tasks. Some experimentation is required to find optimal numbers and types of layers in deep neural networks, but as the theory of deep learning advances, we should have a more solid, formal foundation for reasoning about deep learning network design.

**References**

Ananiadou, Sophia, et al. "Named entity recognition for bacterial type IV secretion systems." *PLoS One*6.3 (2011): e14780.

Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).

Siencnik, Scharolta Katharina. "Adapting word2vec to named entity recognition." *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*. 2015.

Tsuboi, Yuta. "Neural Networks Leverage Corpus-wide Information for Part-of-speech Tagging." *EMNLP*. 2014.

Yin, Wenpeng, and Hinrich Schütze. "Multichannel variable-size convolution for sentence classification." *arXiv preprint arXiv:1603.04513* (2016).