Identification of Adjective-Noun Neologisms using Pretrained Language Models

John P. McCrae
Data Science Institute/Insight Centre for Data Analytics
National University of Ireland Galway
Galway, Ireland
john@mccr.ae

Abstract

Neologism detection is a key task in the constructing of lexical resources and has wider implications for NLP, however the identification of multiword neologisms has received little attention. In this paper, we show that we can effectively identify the distinction between compositional and non-compositional adjective-noun pairs by using pretrained language models and comparing this with individual word embeddings. Our results show that the use of these models significantly improves over baseline linguistic features, however the combination with linguistic features still further improves the results, suggesting the strength of a hybrid approach.

1 Introduction

In the context of the construction of lexical resources, such as WordNet (Miller, 1995; Fellbaum, 2012), a key task is the identifications of terms that would be of relevance for inclusion in the resource and this task is called 'neologism detection.' Detection of single word neologisms can be principally accomplished by means of frequency statistics (McCrae et al., 2017) and even new senses of words can be identified by means of topic models (Lau et al., 2012). However, this task is much harder when we consider multiword expressions as a multiword expression may consist of two or more words that are already in the dictionary but whose combination may give extra meaning that could not be understood from just the words that compose this multiword expression. For example a 'common viper' is not merely a viper that is 'common', but in fact refers to *Vipera berus* a specific species of snake. In contrast, a 'dangerous viper' is simply a viper that is also dangerous and as such most lexicographers would prefer not to include the term in their resources.

In this work, we focus on a particular kind of construction of neologisms, that is neologisms where the term consists of a single adjective and a noun. The reason for this focus is driven by the idea that the semantics of adjectives is complex in terms of their semantic compositionality (McCrae et al., 2014) and this can be broadly broken down into three categories, *intersective*, *subsective* and *privative* adjectives (Partee, 2003; Bouillon and Viegas, 1999; Morzycki, 2015). We use WordNet as the principle background knowledge and thus rely on the judgement of the WordNet lexicographers in order to deduce if a particular adjective-noun combination is a neologism.

Our approach for detecting whether adjective-noun pairs are likely to be neological is based on the recent breakthroughs regarding pretrained language models, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), which have shown to be effective for solving a wide variety of tasks (Radford et al., 2018). For this particular problem of neologism detection, it is clear that there is significant value in the use of these pretrained models as they easily create a vector that represents the adjective-noun combination and this can be compared with a word-based model such as GloVe (Pennington et al., 2014), to deduce if an adjective-noun pair is compositional or neological.

The paper is structured as follows, first in Section 2 we will describe some of the related work in the identification of neologisms, terminology and semantic compositionality. We will then, in Section 3, describe how we created a dataset for noun-adjective neologisms and in particular how we constructed a weak negative set for evaluation. We then describe our baseline methodologies and how we used pretrained language models in order to identify adjective-noun neologism with increased accuracy. The results

Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 135–141
Florence, Italy, August 2, 2019.

of these experiments are presented in Section 4 before we conclude in Section 5. The code and datasets used in these experiments are available at https://github.com/jmccrae/adj-noun-neologism-identification.

2 Related Work

Neologism identification is a task that is a basic task as part of the construction of a lexicon and as the task of lexicography is being increasingly automated (Kosem et al., 2013) in the context of infrastructures such as ELEXIS (Krek et al., 2018), and as such it is of increasing importance. However, while the task has received some attention, most approaches so far have significant weaknesses, even though it is a major area of work for publishers in lexicography (O'Donovan and O'Neill, 2008). Some semi-automated approaches have relied on the extraction of features and the use of classifiers such as SVMs (Falk et al., 2014) or on language-specific features (Breen, 2010).

Of close relationship to this task is automatic term recognition, where new terms are recognized based on their occurrence in a corpus. In these works, a number of metric for assessing 'termhood' (Spasić et al., 2013; Cram and Daille, 2016) have been introduced and these are often developed to work in specific domains (Buitelaar et al., 2013). It has been shown that combinations of many metrics can effectively learn terms (Astrakhantsev, 2014). However, previous work (McCrae et al., 2017) as well as the results in this paper show that these metrics perform poorly at identifying semantic compositionality.

The semantics of adjectives have been studied not only from a logical perspective but as in terms of vector space models and word embeddings and in the context of analysis of semantic compositionality (Mitchell and Lapata, 2008). Most works start from Mitchell and Lapata in representing the compositional vector of an adjective-noun pair with the following equation

$$\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v}$$

Where $\mathbf{p}$ is the vector of compound, $\mathbf{u}$ and $\mathbf{v}$ are vectors for the individual words and α,β are learned weights. This has been extended by replacing the scalar values, α and β with matrices (Boleda et al., 2013):

$$\mathbf{p} = \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v}$$

Dataset	Positives	Negatives
Training	9,474	84,934
Development	1,000	1,000
Test	1,000	1,000
Total	11,474	86,934

Table 1: The number of positive and (weak) negative examples of adjective-nouns used in this study

Further, it has been suggested that adjectives themselves should be matrices (Baroni and Zamparelli, 2010), such that

$$\mathbf{p} = \mathbf{A}_u\mathbf{v}$$

However, learning a matrix to represent each word can be quite difficult. This has been further extended to an approach where each word has a matrix to give a general approach to semantic compositionality (Socher et al., 2012). Moreover, it was shown that simpler models such as bidirectional LSTMs produce better results (Tai et al., 2015). This has lead to the development of pretrained models (Devlin et al., 2018; Peters et al., 2018), which can be trained on truly massive corpora and then still be effectively applied to tasks with relatively little training data.

3 Methodology

3.1 Data Preparation

In order to develop a classifier to determine if a particular adjective-noun pair is a neologism. We first need to develop a set of pairs that we know to be neological and a set that we can assume is likely not to be. For the development of the positive set, we simply took all the two-word expressions within Princeton WordNet 3.1, and deduced the likely part-of-speech tagging using NLTK (Loper and Bird, 2002) and selected only those that were tagged as "`JJ NN`" or "`JJ NNS`". This yielded as set of 11,474 terms that we could use as a positive set.

Developing a negative set is much harder, as we would need to ask an expert lexicographer to manually evaluate a large number of adjective-noun combinations and verify that they were not neologisms that could be put into a dictionary. As such, we rely on a weakly supervised dataset that was constructed from Wikipedia. In particular, we randomly chose from Wikipedia articles a list

of unique adjective-noun pairs, which again were identified by part-of-speech tagging with NLTK, and then filtered out all those pairs, which are already in Wordnet. As this negative set is still likely to contain some true neologisms, we performed a quick manual analysis of 100 of these terms showed that 5 of them were certainly worthy of inclusion in a dictionary (e.g., 'special education', 'safe position') as they have meanings that are not deducible from the two words that compose the phrase. In contrast, most of the examples in the set were clearly compositional, e.g., 'British soldiers', 'much teamwork', 'new congregation'. One example was unclear 'Korean language', which does not occur in WordNet, while other similar terms, such as 'English language' and 'German language' do. As such we estimate that our weak negative set is about 94-95% negative. We acknowledge that this is a weakness of our approach however it would be very expensive to construct a true gold standard and our experiments and analysis below show that the system is capable of effectively learning this task in spite of the noisy training data.

In this way, we constructed a set of weak negative examples that was roughly ten times larger than the positive set, as our intuition was that there are many more negative examples in text than occur naturally. We reserved two sets of 1,000 positive and negative examples for test and development as shown in Table 1.

3.2 Baseline Models

A natural approach for determining whether an adjective-noun pair is compositional would be to compare the frequency with which the adjective-noun occurs in comparison to the adjective and noun's total frequency. This can be achieved by means of Probabilistic Mutual Information as follows:

$$PMI(uv) = p(uv) \log \left(\frac{p(uv)}{p(u)p(v)} \right)$$

Where $p(uv)$ represents the probability of the adjective-noun pair, uv, occurring in our corpus, i.e., the total frequency divided by the length of the corpus, and $p(u)$ and $p(v)$ representing the probability of the adjective, u and the noun v. For corpora we used a recent dump[1] of Wikipedia and we developed this into a simple classifier by learning a threshold, β from the development dataset accepting a pair as a neologism if

$$PMI(uv) > \beta$$

The results from this (in line with our previous experience in this task) were little better than a majority class baseline and as such we developed a classifier that looked only at the words that are in the compound and deduced whether they were neological based on the words themselves. The principal reason for this is that we are attempting to distinguish between collocations and phrases representing novel concepts and it the frequency of these are very similar, meaning that PMI does a very poor job in distinguishing these two similar but distinct linguistic phenomena. In this case we used a *naïve Bayes* classifier which predicts if a word pair is a neologism based on whether $p(\text{Neologism}|uv) > p(\neg\text{Neologism}|uv)$ where:

$$p(\text{Neologism}|uv) \propto$$
$$p(u|\text{Neologism})p(v|\text{Neologism})p(\text{Neologism})$$

The relevant probabilities $p(u|\text{Neologism})$ was simply deduced by the frequency with which a given adjective or noun occurred in our positive or negative training set. The resulting Naïve Bayes classifier provided (surprisingly) strong results and so we continued to use it as a feature within our complete model.

3.3 Using Pretrained Models

We used three pretrained models for computing a single representation of adjective-nouns:

USE Universal sentence encoders (Cer et al., 2018) were introduced to provide a way to make embeddings of whole sentences. As such, they directly model semantic compositionality and we apply them by considering our term as a sentence and generating an 512-dimensional embedding of the term.

ELMo ELMo is a pretrained language model that provides a deep contextual representation of a sentence. We used the 'small' model which generates a representation of 1,024 dimensions.

BERT BERT has further innovated on the pretrained model by training in both direction. We use the final sentence encoding of our

[1] This corpus was compiled in December 2015

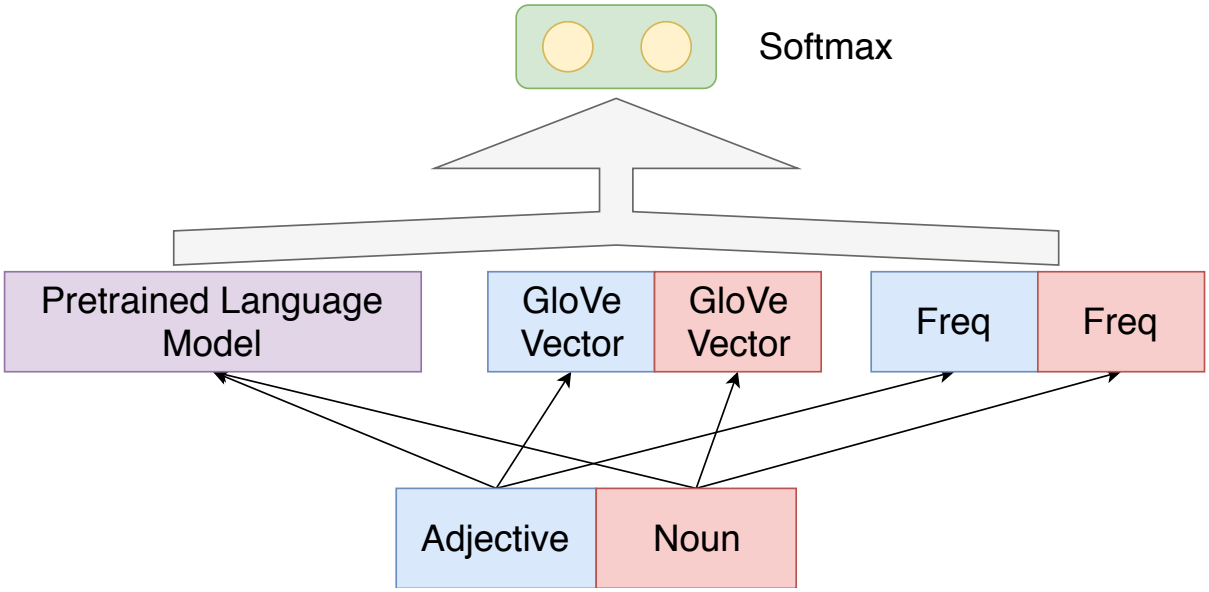

Figure 1: The architecture of the neural network used to identify adjective-noun neologisms

noun-adjective pair, which is a vector of dimensionality 768.

In order to deduce whether there was a significant improvement in the compositional representation that was learnt by these models in contrast to the individual words, we also used a pretrained model for the individual words, namely GloVe (Pennington et al., 2014), which we chose at is has been shown to have good performance across a wide number of tasks. We developed a single vector to represent the noun-adjective by concatenating the two vectors we have from GloVe:

$$\mathbf{g}_{uv} = \begin{pmatrix} \mathbf{g}_u \\ \mathbf{g}_v \end{pmatrix}$$

As we discovered that the Naïve Bayes baseline model was very strong we also calculated for each of the examples the following feature vector:

$$\mathbf{f}_{uv} = \begin{pmatrix} \log(p(u|\text{Neologism})) \\ \log(p(u|\neg\text{Neologism})) \\ \log(p(v|\text{Neologism})) \\ \log(p(v|\neg\text{Neologism})) \end{pmatrix}$$

We combined all these vectors as follows:

$$\mathbf{x} = \mathbf{A}\mathbf{p}_{uv} + \mathbf{B}\mathbf{g}_{uv} + \mathbf{C}\mathbf{f}_{uv} \quad (1)$$

Where $\mathbf{x} \in \mathbb{R}^2$ and we then used a single dense layer taking $\mathbf{x}$ as input to compare the pretrained representation, $\mathbf{p}_{uv}$ with the GloVe representation, $\mathbf{g}_{uv}$. This model is depicted in Figure 1. The error function for the network was cross-entropy over the softmax of the values for $\mathbf{x}$. The softmax was chosen to output two values which represent the probability of a term being neological and not being neological respectively. All models were trained with the Adam optimizer (Kingma and Ba, 2014) for a total of 200 epochs with a learning rate of 0.01 and at the end of each epoch the accuracy on the development set was evaluated and the final model selected for evaluation on the test set was the model with highest development accuracy. In general, this model occurred within the first 100 epochs so we do not expect that more training would lead to better accuracy.

4 Results

We evaluated the model given in Equation 1 in a number of settings, by varying the inclusion of the features from the model. Firstly we considered the model without the use of pretrained language models and only the GloVe vectors which we term the "feed forward" model, this can be considered as fixing the corresponding matrix ($\mathbf{A}$) to zero. We used the GloVe vectors trained on the 6 billion word corpus which comes in four dimensions, 50, 100, 200, 300. We evaluated on all of these settings and in addition the case where we did not use any vectors of GloVe which we labelled as "n/a". As such the setting "feed forward (n/a)" could be considered as another baseline that does not use any features from deep neural networks. We then

Model	GloVe Dimensions	Accuracy	Precision	Recall	F-Measure
PMI (Baseline)	n/a	0.491	0.495	0.979	0.658
Naïve Bayes (Baseline)	n/a	0.800	0.735	**0.937**	0.824
Feed Forward	n/a	0.834	0.850	0.810	0.829
Feed Forward	50	0.846	0.857	0.831	0.844*
Feed Forward	100	0.846	0.818	0.889	0.852†
Feed Forward	200	0.835	0.852	0.810	0.830
Feed Forward	300	0.846	0.854	0.833	0.844*
USE	n/a	0.833	0.869	0.784	0.824
USE	50	0.861	0.852	0.872	0.862†
USE	100	0.873	0.861	0.888	0.874†
USE	200	0.859	0.849	0.872	0.860†
USE	300	0.862	0.844	0.887	0.865†
ELMo	n/a	0.853	0.865	0.836	0.850†
ELMo	50	0.858	0.848	0.872	0.860†
ELMo	100	0.860	0.873	0.841	0.857†
ELMo	200	0.866	0.853	0.884	0.868†
ELMo	300	0.860	0.881	0.832	0.856†
BERT	n/a	0.830	0.808	0.866	0.835
BERT	50	0.862	0.839	0.894	0.866†
BERT	100	**0.882**	**0.895**	0.866	**0.880**†
BERT	200	0.854	0.872	0.830	0.850†
BERT	300	0.848	0.828	0.879	0.853†
BERT (No Freq)	100	0.846	0.834	0.863	0.848†

Table 2: Result for the detection of neological adjective-noun terms using our models. * and † denote a statistically significant improvement over the Naïve Bayes baseline at $p = 0.05, 0.01$ respectively.

evaluated all these settings on the 3 pretrained language models, USE, ELMo and BERT and the results are presented in Table 2. Statistical significance was calculated at two levels (Yeh, 2000).

The strongest result in accuracy, precision and F-Measure is the BERT model with GloVe vectors of dimensionality 100, although the USE and ELMo methods present a similar result with GloVe dimensionality of 100 or 200, suggesting that the use of pretrained models in general is helpful for the identification of neological adjective-noun phrases. The difference in performance between the choice of models was however not statistically significant. Furthermore, we also observe that the larger GloVe vectors are not helpful and observations of the test set accuracy as well as preliminary experiments in more complex neural network architectures have suggested that over-fitting is likely the cause of this given the comparatively small training set.

We found that the inclusion of the frequency feature remained helpful and to evaluate this we rerun our best scoring model with the frequency features and presented them on the bottom row of Table 2, we see that the results without frequency features is still significantly better than the baseline, however the inclusion of these features does give a sizeable increase in the performance of the system. As such, this suggests that there is still a role for traditional feature engineering approaches alongside deep learning methodologies for this task.

Further, we applied a qualitative analysis of the errors made by the system, and we show an example of some of the errors generated by the ELMo-based system in Table 3. For most results it is hard to see why the system made an error, however there are a few patterns, in that many of the false negatives seem to contain low-frequency adjectives such as ‘antigenic’ or ‘Sullian’. In the false

positives, as expected we see some that should not be counted as errors, in particular 'alpha interferon', and this is due to the weaknesses in our methodology that we have previously noted. We also see many cases that would also be hard for a human to decide if they are truly compositional such as 'natural world', 'Korean language' or 'constitutional law', confirming our results that the system is producing near-human results for this task.

5 Conclusion

We have presented a method for identifying adjective-noun pairs as neologisms and have shown that the usage of pretrained language models improves significantly over other baselines. This is particularly interesting as the systems presented in this paper do not require the usage of a large corpus and as such can be robustly and easily applied to a large number of domains. However, we discovered that simple frequency features are still important and this suggests that the combination of linguistically motivated features as well as deep learning models is likely to provide the best results.

Acknowledgments

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund, and the European Unions Horizon 2020 research and innovation programme under grant agreement No 731015, ELEXIS - European Lexical Infrastructure.

References

Nikita Astrakhantsev. 2014. Automatic term acquisition from domain-specific text collection by using Wikipedia. *Proceedings of the institute for system programming*, 26(4):7–20.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. Association for Computational Linguistics.

Gemma Boleda, Marco Baroni, Louise McNally, et al. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 35–46.

Pierrette Bouillon and Evelyne Viegas. 1999. The description of adjectives for natural language processing: Theoretical and applied perspectives. In *Proceedings of Description des Adjectifs pour les Traitements Informatiques. Traitement Automatique des Langues Naturelles*, pages 20–30.

James Breen. 2010. Identification of neologisms in Japanese by corpus analysis. *E-lexicography in the 21st Century: New Challenges, New Applications: Proceedings of ELex 2009, Louvain-la Neuve*, pages 13–21.

Paul Buitelaar, Georgeta Bordea, and Tamara Polajnar. 2013. Domain-independent term extraction through domain modelling. In *The 10th international conference on terminology and artificial intelligence (TIA 2013), Paris, France*. 10th International Conference on Terminology and Artificial Intelligence.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Damien Cram and Béatrice Daille. 2016. Terminology extraction with term variant detection. *Proceedings of ACL-2016 System Demonstrations*, pages 13–18.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ingrid Falk, Delphine Bernhard, and Christophe Gérard. 2014. From non word to new word: Automatically identifying neologisms in French newspapers. In *LREC-The 9th edition of the Language Resources and Evaluation Conference*.

Christiane Fellbaum. 2012. Wordnet. *The Encyclopedia of Applied Linguistics*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Iztok Kosem, Polona Gantar, and Simon Krek. 2013. Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. *Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex*, pages 17–19.

Simon Krek, John McCrae, Iztok Kosem, Tanja Wissek, Carole Tiberius, Roberto Navigli, and Bolette Sandford Pedersen. 2018. European Lexicographic Infrastructure (ELEXIS). In *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts*, pages 881–892.

False Negatives	False positives
Suillus albivelatus	uniform button
critical mass	natural world
Norwegian elkhound	constitutional law
free people	single tube
evolutionary trend	religious knowledge
financial backing	transitional phase
total depravity	pilot error
fluorescent fixture	Korean language
right hand	alpha interferon
antigenic determinant	regulatory region

Table 3: Some examples of false negatives and false positives generated by the system

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.

John P. McCrae, Christina Unger, Francesca Quattri, and Philipp Cimiano. 2014. Modelling the Semantics of Adjectives in the Ontology-Lexicon Interface. In *Proceedings of 4th Workshop on Cognitive Aspects of the Lexicon*.

John P. McCrae, Ian Wood, and Amanda Hicks. 2017. The Colloquial WordNet: Extending Princeton WordNet with Neologisms. In *Proceedings of the First Conference on Language, Data and Knowledge (LDK2017)*, pages 194–202.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, pages 236–244.

Marcin Morzycki. 2015. *The lexical semantics of adjectives: more than just scales*, Key Topics in Semantics and Pragmatics, pages 13–87. Cambridge University Press.

Ruth O'Donovan and Mary O'Neill. 2008. A systematic approach to the selection of neologisms for inclusion in a large monolingual dictionary. In *Proceedings of the 13th Euralex International Congress*, pages 571–579.

Barbara H Partee. 2003. Are there privative adjectives. In *Conference on the Philosophy of Terry Parsons, University of Massachusetts, Amherst*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Self-published.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. 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*, pages 1201–1211. Association for Computational Linguistics.

Irena Spasić, Mark Greenwood, Alun Preece, Nick Francis, and Glyn Elwyn. 2013. Flexiterm: a flexible term recognition method. *Journal of biomedical semantics*, 4(1):27.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 947–953. Association for Computational Linguistics.