# Investigating the Contribution of Distributional Semantic Information for Dialogue Act Classification

**Dmitrijs Milajevs**
Queen Mary University of London
d.milajevs@qmul.ac.uk

**Matthew Purver**
Queen Mary University of London
m.purver@qmul.ac.uk

## Abstract

This paper presents a series of experiments in applying compositional distributional semantic models to dialogue act classification. In contrast to the widely used bag-of-words approach, we build the meaning of an utterance from its parts by composing the distributional word vectors using vector addition and multiplication. We investigate the contribution of word sequence, dialogue act sequence, and distributional information to the performance, and compare with the current state of the art approaches. Our experiment suggests that that distributional information is useful for dialogue act tagging but that simple models of compositionality fail to capture crucial information from word and utterance sequence; more advanced approaches (e.g. sequence- or grammar-driven, such as categorical, word vector composition) are required.

## 1 Introduction

One of the fundamental tasks in automatic dialogue processing is dialogue act tagging: labelling each utterance with a tag relating to its function in the dialogue and effect on the emerging context: *greeting*, *query*, *statement* etc (see e.g. (Core, 1998)). Although factors such as intonation also play a role (see e.g. (Jurafsky et al., 1998)), one of the most important sources of information in this task is the semantic meaning of an utterance, and this is reflected in the fact that people use similar words when they perform similar utterance acts. For example, utterances which state opinion (tagged `sv` in the standard DAMSL schema, see below) often include words such as "*I think*", "*I believe*", "*I guess*" etc. Hence, a similarity-based model of meaning — for instance, a distributional semantic model — should provide benefits over a purely word-based model for dialogue act tagging. However, since utterances generally consist of more than one word, one has to be able to extend such similarity-based models from single words to sentences and/or complete utterances. Hence, we consider here the application of compositional distributional semantics for this task.

Here, we extend bag-of-word models common in previous approaches (Serafin et al., 2003) with simple compositional distributional operations (Mitchell and Lapata, 2008) and examine the improvements gained. These improvements suggest that distributional information does improve performance, but that more sophisticated compositional operations such as matrix multiplication (Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011) should provide further benefits.

The state of the art is a supervised method based on Recurrent Convolutional Neural Networks (Kalchbrenner and Blunsom, 2013). This method learns both the sentence model and the discourse model from the same training corpus, making it hard to understand how much of the contribution comes from the inclusion of distributional word meaning, and how much from learning patterns specific to the corpus at hand. Here, in contrast, we use an external unlabeled resource to obtain a model of word meaning, composing words to obtain representations for utterances, and rely on training data only for discourse learning for the tagging task itself.

We proceed as follows. First, we discuss related work by introducing distributional semantics and describe common approaches for dialogue act tagging in Section 2. Section 3 proposes several models for utterance representation based on the bag of words approach and word vector composition. We describe the experiment and discuss the result in Section 4. Finally, Section 5 concludes the work.

*Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) @ EACL 2014*, pages 40–47,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

## 2 Related work

**Distributional semantics** The aim of natural language semantics is to provide logical representations of meaning for information in textual form. Distributional semantics is based on the idea that "You shall know a word by the company it keeps" (Firth, 1957) – in other words, the meaning of a word is related to the contexts it appears in. Following this idea, word meaning can be represented as a vector where its dimensions correspond to the usage contexts, usually other words observed to co-occur, and the values are the co-occurrence frequencies. Such a meaning representation is easy to build from raw data and does not need rich annotation.

Methods based on this distributional hypothesis have recently been applied to many tasks, but mostly at the word level: for instance, word sense disambiguation (Zhitomirsky-Geffet and Dagan, 2009) and lexical substitution (Thater et al., 2010). They exploit the notion of similarity which correlates with the angle between word vectors (Turney et al., 2010). *Compositional* distributional semantics goes beyond the word level and models the meaning of phrases or sentences based on their parts. Mitchell and Lapata (2008) perform composition of word vectors using vector addition and multiplication operations. The limitation of this approach is the operator associativity, which ignores the argument order, and thus word order. As a result, "*John loves Mary*" and "*Mary loves John*" get assigned the same meaning.

To capture word order, various approaches have been proposed. Grefenstette and Sadrzadeh (2011) extend the compositional approach by using non-associative linear algebra operators as proposed in the theoretical work of (Coecke et al., 2010). Socher et al. (2012) present a recursive technique to build compositional meaning of phrases from their constituents, where the nonlinear composition operators are learned by Neural Networks.

**Dialogue act tagging** There are many ways to approach the task of dialogue act tagging (Stolcke et al., 2000). The most successful approaches combine *intra*-utterance features, such as the (sequences of) words and intonational contours used, together with *inter*-utterance features, such as the sequence of utterance tags being used previously. To capture both of these aspects, sequence models such as Hidden Markov Models are widely used (Stolcke et al., 2000; Surendran and Levow, 2006). The sequence of words is an observable variable, while the sequence of dialogue act tags is a hidden variable.

However, some approaches have shown competitive results without exploiting features of inter-utterance context. Webb et al. (2005) concentrate only on features found inside an utterance, identifying ngrams that correlate strongly with particular utterance tags, and propose a statistical model for prediction which produces close to the state of the art results.

The current state of the art (Kalchbrenner and Blunsom, 2013) uses Recurrent Convolutional Neural Networks to achieve high accuracy. This model includes information about word identity, intra-utterance word sequence, and inter-utterance tag sequence, by using a vector space model of words with a compositional approach. The words vectors are not based on distributional frequencies in this case, however, but on randomly initialised vectors, with the model trained on a specific corpus. This raises several questions: what is the contribution of word sequence and/or utterance (tag) sequence; and might further gains be made by exploiting the distributional hypothesis?

As our baseline, we start with an approach which uses only word information, and excludes word sequence, tag sequence and word distributions. Serafin et al. (2003) use Latent Semantic Analysis for dialogue act tagging: utterances are represented using a bag-of-words representation in a word-document matrix. The rows in the matrix correspond to words, the columns correspond to documents and each cell in the matrix contains the number of times a word occurs in a document. Singular Value Decomposition (SVD) is then applied to reduce the number of rows in the matrix, with the number of components in the reduced space set to 50. To predict the tag of an unseen utterance, the utterance vector is mapped to the reduced space and the tag of the closest neighbor is assigned to it (using cosine similarity as a similarity measure). The reported accuracy on the Spanish Call Home corpus for predicting 37 different utterance tags is 65.36%.

## 3 Utterance models

In this paper, we investigate the extent to which distributional representations, word order information, and utterance order information can improve this basic model, by choosing different ways to represent an utterance in a vector space. We design three basic models. The first model is based directly on the bag-of-words model which serves as the baseline in our experiment, following (Serafin et al., 2003); and extends this to investigate the effect of word order information by moving from word unigrams to bigrams. The second model investigates distributional information, by calculating word vector representations from a general corpus, and obtaining utterance representations by composing the word vectors using simple operators. The third model extends this idea to investigate the role of utterance order information, by including the information about the previous utterance.

**Bag of words** The first model represents an utterance as a vector where each component corresponds to a word. The values of vector components are the number of times the corresponding words occured in the utterance. The model is similar to (Serafin et al., 2003), but the matrix is transposed. We refer to it as *bag of unigrams* in Table 1.

However, this bag of words approach does not preserve any word order information. As it has been said previously, for the dialogue act tagging word order may be crucial. Consider these utterances:

- *John, are there cookies*
- *John, there are cookies*

One of the utterances is a question (or request) while the other is a statement. However, the bag of words model will extract the same vector representation for both.

To overcome this problem we also represent an utterance as a *bag of bigrams*. When bigrams are used in place of single words, the utterance representation will differ. The question contains the bigram "*are there*", while the statement contains the bigram "*there are*".

**Simple composition** Our second model exploits the distributional hypothesis, by representing words not as atomic types (i.e. individual dimensions in the utterance matrix, as above), but as vectors encoding their observed co-occurrence distributions. We estimate these from a large corpus of general written English (the Google Books Ngrams corpus – see below).

However, this raises the question of how to compose these word vectors into a single representation for an utterance. Various approaches to compositional vector space modelling have been successfully applied to capture the meaning of a phrase in a range of tasks (Mitchell and Lapata, 2008; Grefenstette and Sadrzadeh, 2011; Socher et al., 2013). In this work, we follow (Mitchell and Lapata, 2008) and apply vector addition and pointwise multiplication to obtain the vector of an utterance from the words it consists of. This has the advantage of simplicity and domain-generality, requiring no sentence grammar (problematic for the non-canonical language in dialogue) or training on a specific corpus to obtain the appropriate compositionality operators or associative model; but has the disadvantage of losing word order information. The corresponding models are referred as *addition* and *multiplication* in Table 1 and Table 2.

**Previous utterance** A conversation is a sequence of utterances, and the tag of an utterance often depends on the previous utterance (e.g. answers tend to follow questions). Hidden Markov Models (Surendran and Levow, 2006; Stolcke et al., 2000) are often used to capture these dependencies; Recurrent Convolutional Neural Networks (Kalchbrenner and Blunsom, 2013) have been used to simultaneously capture the intra-utterance sequence of words and the inter-utterance sequence of dialog tags in a conversation.

In this model, we are interested specifically in the effect of inter-utterance (tag) sequence. We provide *previous addition* and *previous multiplication* models as simple attempts to capture this phenomenon: the vector of an utterance is the concatenation of its vector obtained in the corresponding compositional model (*addition* or *multiplication*) and the vector of the previous utterance.

## 4 Predicting dialogue acts

### 4.1 The resources

This section describes the resources we use to evaluate and compare the proposed models.

**Switchboard corpus** The Switchboard corpus (Godfrey et al., 1992) is a corpus of telephone conversations on selected topics. It consists of about 2500 conversations by 500 speakers from the U.S. The conversations in the corpus are labeled with 42 unique dialogue act tags and split to 1115 train

```
A o    : Okay. /
A qw   : {D So, }
B qy^d: [ [I guess, +
A +    : What kind of experience
         [ do you, + do you ] have,
          then with child care? /
B +    : I think, ] + {F uh, }
         I wonder if that worked. /
```

(a) A conversation with interrupted utterances.

```
A o    : Okay.
A qw   : So What kind of experience
         do you do you have then
         with child care?
B qy^d: I guess I think uh I wonder
         if that worked.
```

(b) A preprocessed conversation.

Figure 1: A example of interrupted utterances from Switchboard and their transformation.

and 19 test conversations (Jurafsky et al., 1997; Stolcke et al., 2000).

In addition to the dialog act tags, utterances interrupted by the other speaker (and thus split into two or more parts) have their continuations marked with a special tag "+". Tag prediction of one part of an interrupted utterance in isolation is a difficult task even for a human; for example, it would not be clear why the utterance *"So,"* should be assigned the tag `qw` (wh-question) in Figure 1a without the second part *"What kind of experience do you have [. . . ]"*. Following (Webb et al., 2005) we preprocess Switchboard by concatenating the parts of an interrupted utterance together, giving the result the tag of the first part and putting it in its place in the conversation sequence. We also remove commas and disfluency markers from the raw text. Figure 1b illustrates the transformation we do as preprocessing.

We split the utterances between training and testing as suggested in (Stolcke et al., 2000).

**Google Books Ngram Corpus** The Google Books Ngram Corpus (Lin et al., 2012) is a collection of n-gram frequencies over books written in 8 languages. The English part of the corpus is based on more than 4.5 million books and contains more than four thousand billion tokens. The resource provides frequencies of n-grams of length 1 to 5. For our experiments we use 5-grams from the English part of the resource.

### 4.2 Word vector spaces

In distributional semantics, the meanings of words are captured by a vector space model based on a word co-occurrence matrix. Each row in the matrix represents a target word, and each column represents a context word; each element in the matrix is the number of times a target word co-occured with a corresponding context word. The frequency counts are typically normalized, or weighted using tf-idf or log-likelihood ratio to obtain better results, see (Mitchell and Lapata, 2008; Agirre et al., 2009) for various approaches. It is also common to apply dimensionality reduction to get higher performance (Dinu and Lapata, 2010; Baroni and Zamparelli, 2010).

As target words we select all the words in our (Switchboard) training split. As context words we choose the 3000 most frequent words in the Google Ngram Corpus, excluding the 100 most frequent. To obtain co-occurrence frequencies from ngrams we sum up the frequency of a 5-gram over the years, treat the word in the middle as a target, and the other words as its contexts.

For normalization, we experiment with a vector space based on raw co-occurrences; a vector space where frequencies are weighted using tf-idf; and another one with the number of dimensions reduced to 1000 using Non-negative Matrix Factorization (NMF) (Hoyer, 2004).

We use the NMF and tf-idf implementations provided by `scikit-learn` version 0.14 (Pedregosa et al., 2011). For tf-idf, the term vectors are $L^2$ normalized. For NMF, NNDSVD initialization (Boutsidis and Gallopoulos, 2008) is used, and the tolerance value for stopping conditions is set to 0.001. The co-occurrence matrix is line-normalized, so the sum of the values in each row is 1 before applying NMF.[1]

### 4.3 Evaluation

To evaluate these possible models we follow (Serafin et al., 2003). Once we have applied a model to extract features from utterances and build a vector space, the dimensionality of the vector space is reduced using SVD to 50 dimensions. Then a k-nearest neighbours (KNN) classifier is trained and used for utterance tag prediction. In contrast to (Serafin et al., 2003), we use Euclidean distance as a distance metric and choose the most

[1] The co-occurrence matrix and the information about the software used in the experiment are available at http://www.eecs.qmul.ac.uk/~dm303/cvsc14.html

| Method | Accuracy |
|---|---|
| (Kalchbrenner and Blunsom, 2013) | **0.739** |
| (Webb et al., 2005) | 0.719 |
| (Stolcke et al., 2000) | 0.710 |
| (Serafin et al., 2003) | *0.654* |
| Bag of unigrams | 0.602 |
| Bag of bigrams | 0.621 |
| Addition | 0.639 |
| Multiplication | 0.572 |
| Previous addition | 0.569 |
| Previous multiplication | 0.497 |

Table 1: Comparison with previous work. Note that (Serafin et al., 2003) do not use Switchboard and therefore their results are not directly comparable to others.

| | Space | | |
|---|---|---|---|
| Model | Raw | tf-idf | NMF |
| Addition without SVD | 0.592 | | |
| Addition | 0.610 | **0.639** | 0.620 |
| Multiplication | 0.572 | 0.568 | 0.525 |
| Previous addition | | | 0.569 |
| Previous multiplication | | | 0.497 |

Table 2: Accuracy results for different compositional models and vector spaces.

frequent label among the 5 closest neighbors. The SVD and KNN classifier implementations in `scikit-learn` are used.

**Baseline** In our experiments, the bag of unigrams model accuracy of 0.602 is lower than the accuracy of 0.654 reported in (Serafin et al., 2003), see Table 1. The lower performance may be due to the differences between Switchboard and CallHome37 corpora, in particular the tag distribution.[2] In CallHome37, 42.7% of utterances are labeled with the most frequent dialogue act, while the figure in Switchboard is 31.5%; the more even distribution in Switchboard is likely to make overall average accuracy levels lower.

**Word order** As Table 1 shows, the bag of bigrams model improves over unigrams. This confirms that word order provides important information for predicting dialogue act tags.

**Distributional models** Performance of compositional distributional models depends both on compositional operator and weighting. Table 2 demonstrates accuracy of the models. We instantiate 3 vector spaces from Google Ngrams: one space with raw co-occurrence frequencies, a tf-idf weighted space and a reduced space using NMF.

Addition outperforms multiplication in our experiments, although for other tasks multiplication has been shown to perform better (Grefenstette and Sadrzadeh, 2011; Mitchell and Lapata, 2008). Lower multiplication performance here might be due to the fact that some utterances are rather long (for example, more than 70 tokens), and the resulting vectors get many zero components.

Selection of the optimal weighting method could be crucial for overall model performance. The 3 weighting schemes we use give a broad variety of results; more elaborate weighting and context selection might give higher results.

Figure 2 illustrates dialog tag assignment using addition and the tf-idf weighted vector space. As we do not use any inter-utterance features, the first two statements, which consist only of the word *Okay*, got assigned wrong tags. However, the Wh-question in the conversation got classified as a Yes-No-question, probably because *what* did not influence the classification decision strongly enough and could have been classified correctly using only intra-utterance features. Also, the example shows how important grammatical features are: the verb *think* appears in many different context, and its presence does not indicate a certain type of an utterance.

In addition, we observed that SVD improves classification accuracy. The accuracy of KNN classification without prior dimensionality reduction drops from 0.610 to 0.592 for vector addition on the raw vector space.

**Utterance sequence** To solve the issue of utterances that can be tagged correctly only by considering inter-utterance features, we included previous utterance. However, in our experiment, such inclusion by vector concatenation does not improve tagging accuracy (Table 2). The reason for this could be that after concatenation the dimensionality of the space doubles, and SVD can not handle it properly. We evaluated only dimensionally reduced spaces because of the memory limit.

[2] The CallHome37 corpus is not currently available to us.

```
B **  (b) : Okay.
A b^m (b) : Okay.
B qw  (qy): Well what do you think about the idea of uh kids having to do public
            service work for a year?
B qy  (sd): Do you think it's a <breathing>
A sv  (sv): Well I I think it's a pretty good idea.
A sv  (sd): I think they should either do that or or afford some time to the military
            or or helping elderly people.
B aa  (aa): Yes
B aa  (b) : yes
B %   (%) : def
A sv  (sv): I I you know I think that we have a bunch of elderly folks in the country
            that could use some help
```

Figure 2: The beginning of the conversation 2151 from the test split of Switchboard. In brackets the tags predicted using vector addition as a composition method on the tf-idf space are given. We mark fo_o_fw_"_by_bc as **.

**Summary** Our accuracy is lower compared to other work. Webb et al. (2005)'s method, based only on intra-utterance lexical features, but incorporating longer ngram sequences and feature selection, yields accuracy of 0.719. Advanced treatment of both utterance and discourse level features yields accuracy of 0.739 (Kalchbrenner and Blunsom, 2013). However, our experiments allow us to evaluate the contribution of various kinds of information: vector spaces based on word bigrams and on co-occurrence distributions both outperformed the bag of words approach; but incorporation of previous utterance information did not.

## 5 Conclusions and future work

In this work we evaluated the contribution of word and utterance sequence, and of distributional information using simple compositional vector space models, for dialogue act tagging. Our experiments show that information about intra-utterance word order (ngrams), and information about word co-occurence distributions, outperforms the bag of words models, although not competitive with the state of the art given the simplistic compositional approach used here. Information about utterance tag sequence, on the other hand, did not.

The usage of an external, large scale resource (here, the Google Ngram Corpus) to model word senses improves the tagging accuracy in comparison to the bag of word model, suggesting that the dialogue act tag of an utterance depends on its semantics.

However, the improvements in performance of the *bag of bigrams* model in comparison to *bag of unigrams*, and the much higher results of Webb et al. (2005)'s intra-utterance approach, suggest that the sequence of words inside an utterance is crucial for the dialogue act tagging task. This suggests that our simplistic approaches to vector composition (addition and multiplication) are likely to be insufficient: more advanced, sequence- or grammar-driven composition, such as categorical composition (Coecke et al., 2010), might improve the tagging accuracy.

In addition, our results show that the performance of distributional models depends on many factors, including compositional operator selection and weighting of the initial co-occurrence matrix. Our work leaves much scope for improvements in these factors, including co-occurrence matrix instantiation. For example, the window size of 2, which we used to obtain co-occurrence counts, is lower than the usual size of 5 (Dinu and Lapata, 2010), or the sentence level (Baroni and Zamparelli, 2010). Word representation in a vector space using neural networks might improve accuracy as well (Mikolov et al., 2013).

Previous approaches to dialogue act tagging have shown utterance/tag sequence to be a useful source of information for improved accuracy (Stolcke et al., 2000). We therefore conclude that the lower accuracy we obtained using models that include information about the previous utterance is due again to our simplistic method of composition (vector concatenation); models which reflect dialogue structure or sequence explicitly are likely to be more suited. Kalchbrenner and Blunsom (2013) give one way in which this can be achieved by learning from a specific corpus, and the question of possible alternatives and more general models remains for future research.

## Acknowledgments

We thank Mehrnoosh Sadrzadeh for her helpful advice and valuable discussion. We would like to thank anonymous reviewers for their effective comments. Milajevs is supported by the EPSRC project EP/J002607/1. Purver is supported in part by the European Community's Seventh Framework Programme under grant agreement no 611733 (ConCreTe).

## References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

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.

Christos Boutsidis and Efstratios Gallopoulos. 2008. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *CoRR*, abs/1003.4394.

Mark Core. 1998. Analyzing and predicting patterns of damsl utterance tags. In *Proceedings of the AAAI spring symposium on Applying machine learning to discourse processing*.

G. Dinu and M. Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172. Association for Computational Linguistics.

John R. Firth. 1957. A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis*, pages 1–32.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404. Association for Computational Linguistics.

Patrik O Hoyer. 2004. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder. Institute of Cognitive Science.

Daniel Jurafsky, Elizabeth Shriberg, Barbara Fox, and Traci Curl. 1998. Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of the ACL-COLING Workshop on Discourse Relations and Discourse Markers*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 119–126, Sofia, Bulgaria, August. Association for Computational Linguistics.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Riccardo Serafin, Barbara Di Eugenio, and Michael Glass. 2003. Latent semantic analysis for dialogue act classification. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 94–96. Association for Computational Linguistics.

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.

Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. 2013. Parsing with compositional vector grammars. In *In Proceedings of the ACL conference*. Citeseer.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Carol Van Ess-Dykema, Rachel Martin, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Dinoj Surendran and Gina-Anne Levow. 2006. Dialog act tagging with support vector machines and hidden markov models. In *INTERSPEECH*.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 948–957, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Nick Webb, Mark Hepple, and Yorick Wilks. 2005. Dialogue act classification based on intra-utterance features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding*. Citeseer.

M. Zhitomirsky-Geffet and I. Dagan. 2009. Bootstrapping distributional feature vector quality. *Computational Linguistics*, 35(3):435–461.