Leveraging Preposition Ambiguity to Assess Compositional Distributional Models of Semantics

Samuel Ritter*
Princeton University

Cotie Long
Indiana University

Denis Paperno
University of Trento

Marco Baroni
University of Trento

Matthew Botvinick
Princeton University

Adele Goldberg
Princeton University

Abstract

Complex interactions among the meanings of words are important factors in the function that maps word meanings to phrase meanings. Recently, compositional distributional semantics models (CDSM) have been designed with the goal of emulating these complex interactions; however, experimental results on the effectiveness of CDSM have been difficult to interpret because the current metrics for assessing them do not control for the confound of lexical information. We present a new method for assessing the degree to which CDSM capture semantic interactions that dissociates the influences of lexical and compositional information. We then provide a dataset for performing this type of assessment and use it to evaluate six compositional models using both co-occurrence based and neural language model input vectors. Results show that neural language input vectors are consistently superior to co-occurrence based vectors, that several CDSM capture substantial compositional information, and that, surprisingly, vector addition matches and is in many cases superior to purpose-built paramaterized models.

1 Introduction

Consider the meanings of the following phrases: "red apple," "red hair," and "red state." The meaning of the word "red" in each of these examples interacts with the meaning of the noun it modifies, applying a different color to the first two and a political affiliation to the third. This is an example of a common phenomenon in natural language in which the meaning of a whole expression is not derived from a simple concatenation of its parts, but is composed by interactions among their meanings.

Cognitive and computer scientists have pointed out this complexity and proposed various models for accommodating it (Kintsch, 2001; Mitchell and Lapata, 2010; Socher et al., 2013). A dominant modeling approach seeks to learn functions that combine word representations derived from the distributional structure of large natural language corpora (Deerwester et al., 1990; Landauer and Dumais, 1997). Because the word representations to be combined and the compositional functions are generated based on the distributions of words in corpora, these models have been dubbed compositional distributional semantic models, or CDSM (Marelli et al., 2014). CDSM produce fixed-dimensional vector representations of arbitrary sentences and phrases, and the foundational principle of these models is, stated simply, that semantically similar phrases should have vector representations that are close together in the vector space.

1.1 CDSM Assessment

Past studies have tested how well CDSM adhere to this principle by comparing the vector similarity of pairs of sentences with similarity ratings given by humans. Many of these studies used datasets in which the amount of lexical overlap between the sentence pairs is not carefully controlled, e.g., the datasets of Dolan and Brockett (2005) and Agirre

*Please address correspondence to the first author at swritter@princeton.edu

et al. (2014). One such study obtained the influential result that on such a dataset, simple composition models such as vector addition perform comparably to a state-of-the-art composition model (Blacoe and Lapata, 2012). The success of these simplistic models led to the conjecture that these data sets fail to assess critical aspects of language (Baroni et al., 2014a) and leaves open the question of whether CDSM would outperform simplistic models in a setting in which lexical cues are uninformative.

In the present study, we develop a method for removing the confound of lexical cues from CDSM assessment. The method is to create a set of sentences where each sentence fits into a semantic category and where a sentence's semantic category cannot be determined based on any individual word in the sentence. CDSM are then challenged to create a vector space in which the representations for sentences in a given category cluster together, even though the individual word vectors do not cluster together. This clustering can be tested by training a simple linear classifier on the CDSM representations, then testing it on representations for held out sentences.

Here, we build a suitable test set by leveraging the lexical ambiguity inherent in locative expressions. Locative expressions are phrases that describe a spatial relationship between two objects using two nouns joined by a preposition; for example, "The magnet is on the refrigerator", which describes the relationship of adhesion to a vertical surface. Crucially, the spatial relationship between the two nouns in a locative expression is undetermined by the spatial preposition, and can only be determined based on semantic interactions among the prepositions and the two nouns (Herskovits, 1985).

For example, while "The magnet is on the refrigerator" describes the spatial relationship of adhesion to a vertical surface, "The apple is on the refrigerator" describes support by a horizontal surface. In order to classify a new sentence, e.g., "The magnet is on the papers", into the correct category of support by a horizontal surface, the CDSM vectors for the three sentences must encode the fact that "The magnet is on the papers" shares a common spatial relationship with "The apple is on the refrigerator" and not with "The magnet is on the refrigerator", even though the latter pair of sentences share more words than the former.

Given this dissociation between lexical overlap and spatial relationship, we were able to construct a dataset wherein lexical information is uninformative, and models must rely on compositionality to score well in classification.

1.2 Relation to Past Work

This approach to CDSM assessment is similar to a previous method wherein polysemous verbs are paired with disambiguating nouns in transitive or intransitive verb phrases. These phrases are then matched with "landmark" verbs that are either similar or not similar in meaning to the full phrase. CDSM are then challenged to create representations of the phrases from which classifiers can determine whether or not a phrase is similar to its landmark verb (Kintsch, 2001; Mitchell and Lapata, 2008; Mitchell and Lapata, 2010; Grefenstette and Sadrzadeh, 2011). Another notable CDSM assessment task involves matching a phrase with a word with a similar meaning, for example, matching a short dictionary definition with the word it defines (Kartsaklis et al., 2012; Turney, 2014).

While these methods are applicable only to simple phrases that can be mapped reasonably to a single word, the present method can, in principle, be applied to any type of phrase. This allowed us to build a dataset that extends the current landmark word and word matching datasets in at least two important ways. First, it includes function words, specifically prepositions. Second, it requires the characterization of interactions among three words in each expression, whereas previous datasets had two words per expression, or subsets of the words did not interact in complex ways.

Other important approaches to CDSM assessment include rating the similarity of sentence pairs, determining whether two sentences are paraphrases (Dolan and Brockett, 2005), classifying the entailment relationship between two sentences (Marelli et al., 2014), classifying the relationship between two entities named in a sentence (Hendrickx et al., 2009), and classifying the valence of the sentiment expressed in a sentence (Socher et al., 2013). These methods have primarily been aimed at assessing CDSM on the full array of constructions inherent in naturally generated language, while our method aims to isolate a specific construction of interest.

Category	Example
Adhesion to Vertical Surface	"There is a magnet on the refrigerator."
Support by Horizontal Surface	"There is an apple on the refrigerator."
Support from Above	"There is an apple on the branch."
Full Containment	"There is an apple in the refrigerator."
Partial Containment	"There is an apple in the water."

Table 1: Categories and Example Sentences

2 The Dataset

A list of all of the spatial categories with examples is given in Table 1. The authors chose the set of categories to produce the desired dissociation between lexical meaning and phrase category, taking inspiration from the observations of Herskovits (1985). To produce a dataset of expressions fitting these categories, the first and second authors - both native English speakers - generated a large set of locative expressions, intending each expression for a specific category. Then all of the expressions were independently rated by the first two authors, and any expression for which the ratings disagreed were excluded from the dataset. In order to achieve a balanced category size, the second author then created additional sentences intended for underrepresented categories. All additional sentences were stripped of labels and rated independently by the first author. If the first and second authors' categorizations did not match, the sentence was not added to the dataset.

The dataset contains 500 sentences in total with 100 sentences per category. There is a large amount of lexical variety in the set, with 242 distinct words occurring in noun position one and 213 occurring in noun position two. The dataset is publicly available for download at www.princeton.edu/~swritter.

3 Evaluation Setup

Classification among the five categories was performed using a naive Bayes classifier. Two of the categories contained "in" as the preposition in all sentences while the other three contained "on" in all sentences. To be certain that the held out sentences on which the classifier was tested did not contain even a single category-informative noun, we operationally defined informativeness and relegated all sentences with an informative noun to the training set. A noun was deemed informative if it both occurred more than once in the entire data set and it occurred more frequently in one category than in any other. This criterion yields a set of 80 sentences with no informative nouns, and a set of 420 sentences with at least one informative noun. By this method, we ensure that no component of the models' classification accuracy on the test set is due to the recognition of individual nouns.

In addition to the CDSM, we included two non-distributional models for comparison. The first, referred to as word overlap, consists of a binary feature vector containing one feature per vocabulary item. This model's performance provides an upper-bound on the performance that a model can achieve given only the distribution of word tokens in the training set. The second model, inspired by Srikumar and Roth (2013), contains binary features for Wordnet hypernyms (up to 4 levels) of each sense of the noun and a binary feature for each preposition. This model's score provides an indication of the amount of task-relevant information contained in the taxonomic features of individual words.

We compared CDSM to a further control that consisted of the concatenation of the word vectors. The concatenated vectors contain a complete representation of all of the individual word information, so that any performance the CDSM can achieve above the concatenation score can be attributed to semantic interaction information contained in the parameters of the CDSM.[1]

[1] One other experiment we considered was to test the models on the dataset phrases with prepositions removed. However, LF and PLF are undefined for such an input, and the element-wise models trivially perform better with the preposition included because the preposition is the only word that isn't stripped of informativeness by design of the task. As such, we excluded this experiment from this report.

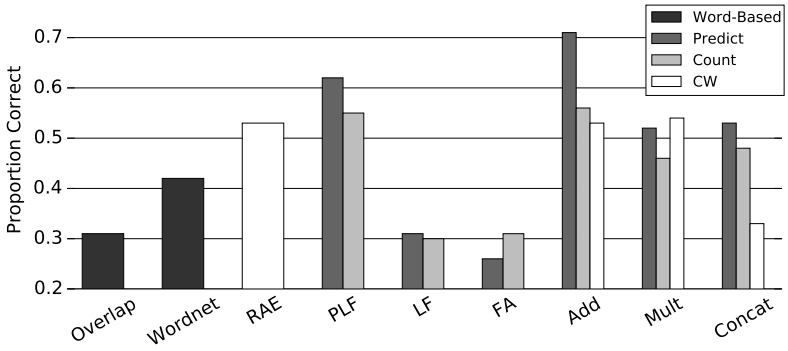

Figure 1: Naive Bayes accuracy scores for count and predict variants of several CDSM. Chance performance on this task was 0.2. Overlap refers to the word overlap baseline. CW refers to the vectors from Collobert and Weston (2008)

.

3.1 Compositional Distributional Models

We compared six models that are currently prominent in the CDSM literature: addition, multiplication (Mitchell and Lapata, 2008), lexical function (LF) (Coecke et al., 2010), practical lexical function (PLF) (Paperno et al., 2014), full additive (FA) (Guevara, 2010; Zanzotto et al., 2010), and the recursive auto-encoder (RAE) (Socher et al., 2011).

The training data for LF, PLF, and FA was the UKWAC+Wikipedia+BNC 2.8 billion word corpus. In training LF, we followed Grefenstette et al. (2013), employing a two-step training regime using corpus-extracted vectors for noun–preposition–noun combinations to estimate matrices of corresponding prepositional phrases, which were in turn used to estimate a three-way tensor of each preposition. For PLF and FA, we learned separate matrices for combining prepositions with each of the two nouns in the construction, using corpus-based vectors of prepositional phrases for training preposition–noun combination. For training composition of the head noun with the prepositional phrase, we used corpus-extracted noun+preposition (for lexical matrices in PLF) or attributive adjective+noun (for attributive construction in FA) vectors. Phrase vectors for training were built as DISSECT 'peripheral' spaces from phrase cooccurrence data in the count models. In the predict models, phrase vectors were learned along with word vectors in one pass, feeding all phrases of the relevant type as single tokens.

The RAE vectors were computed using Socher et al.'s implementation which is trained on a 150K sentence subset of the NYT and AP sections of the Gigaword corpus.

For all compositional models, we used as input two varieties of word level representations: co-occurrence based (Turney et al., 2010) and neural language model (Mikolov et al., 2013). Following Baroni et al. (2014b), we will refer to these variants as *count* and *predict* models respectively. Both word models were trained on the same corpus as those used to train the compositional models. Count was based on a 5 word window weighted with positive PMI and was reduced to 300 dimensions via SVD, while predict was based on a 5 word window using Mikolov's continuous bag of words approach with negative sampling (Mikolov et al., 2013). These parameters were based on their strong performance in the systematic evaluation by Baroni et al. (2014b). Socher et al.'s RAE implementation composes neural language model vectors described by Collobert and Weston (2008) and supplied by Turian et al. (2010). For comparison with the RAE, we report results for addition, multiplication, and concatenation of these same embeddings.

4 Results

The naive Bayes accuracy scores for all models are displayed in Figure 1. Addition, PLF, and the RAE each substantially outperformed concatenation, indicating that these models' vectors contain informa-

tion about the semantic interactions between phrase constituents. Addition scored higher than PLF, while the RAE achieved comparable performance to its additive counterpart. In all cases except FA in which predict and count vectors were compared, predict achieved a higher score. This last result shows that the superiority of predict vectors documented by Baroni et al. (2014b) extends to their use in compositional models.

All of the models performed well above chance accuracy of 0.2. The Wordnet based model achieved accuracy substantially above word overlap using hypernym information, indicating that although each noun is uninformative, its membership in higher level semantic categories is informative. All of the distributional models outperform the non-distributional models, except for LF and FA, which also fail to outperform concatenations of their input vectors. One explanation for the poor performance of LF and FA is that the 2.8B word corpus used to train them did not have sufficient relevant information to specify their large sets of parameters. This explanation is supported by the fact that PLF, a model designed as a parameter-reduced version of LF, performs well.

5 Discussion

The most important finding of this study is that, even on a test painstakingly designed to exclusively assess composition, vector addition matches or outperforms sophisticated CDSM. This finding implies that the structure of distributional vector spaces admits the effective use of addition for modeling complex interactions between meanings. This suggests that future work should be concerned with understanding the properties of distributional vector spaces that make this possible, as well as with understanding how these properties can be leveraged by sophisticated models.

A further contribution of this work is that it serves as a proof-of-concept for a new method for dissociating the influences of lexical and compositional influences on CDSM performance. Future work can extend this approach by finding alternatives to locative expressions in order to test a wider variety of constructions. More immediately, future work may improve the locative expressions dataset by using crowdsourcing to obtain naive participant ratings to corroborate the expert ratings and to increase the size of the dataset.

Acknowledgments

Denis Paperno and Marco Baroni were supported by ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES). Samuel Ritter and Matthew Botvinick were supported by Intelligence Advanced Research Projects Activity (IARPA) Grant n. 102-01.

References

Eneko Agirre, Carmen Baneab, Claire Cardiec, Daniel Cerd, Mona Diabe, Aitor Gonzalez-Agirre, Weiwei Guof, Rada Mihalceab, German Rigaua, and Janyce Wiebeg. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. *SemEval 2014*, page 81.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014a. Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, 9.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014b. Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1.

William Blacoe and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *EMNLP*, pages 546–556.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proc. of IWP*.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *EMNLP*, pages 1394–1404.

Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. 2013. Multi-step regression learning for compositional distributional semantics. In *Proceedings of IWCS*, pages 131–142, Potsdam, Germany.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics*, pages 33–37.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics.

Annette Herskovits. 1985. Semantics and pragmatics of locative expressions*. *Cognitive Science*, 9(3):341–378.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2012. A unified sentence space for categorical distributional-compositional semantics: Theory and experiments. In *In Proceedings of COLING: Posters*. Citeseer.

Walter Kintsch. 2001. Predication. *Cognitive Science*, 25(2):173–202.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *SemEval-2014*.

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 *ACL*, pages 236–244. Citeseer.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Denis Paperno, Nghia The Pham, and Marco Baroni. 2014. A practical and linguistically-motivated approach to compositional distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 90–99, Baltimore, Maryland, June.

Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In *Advances in Neural Information Processing Systems 24*.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642. Citeseer.

V. Srikumar and D. Roth. 2013. Modeling semantic relations expressed by prepositions. In *Transactions of the Association for Computational Linguistics*, volume 1, pages 231–242.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. 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.

Peter D Turney. 2014. Semantic composition and decomposition: From recognition to generation. *arXiv preprint arXiv:1405.7908*.

Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1263–1271.