One Representation per Word — *Does it make Sense for Composition?*

Thomas Kober, Julie Weeds, John Wilkie, Jeremy Reffin and **David Weir**
TAG laboratory, Department of Informatics, University of Sussex
Brighton, BN1 9RH, UK
{t.kober, j.e.weeds, jw478, j.p.reffin, d.j.weir}@sussex.ac.uk

Abstract

In this paper, we investigate whether an *a priori* disambiguation of word senses is strictly necessary or whether the meaning of a word in context can be disambiguated through composition alone. We evaluate the performance of off-the-shelf single-vector and multi-sense vector models on a benchmark phrase similarity task and a novel task for word-sense discrimination. We find that single-sense vector models perform as well or better than multi-sense vector models despite arguably less clean elementary representations. Our findings furthermore show that simple composition functions such as pointwise addition are able to recover sense specific information from a single-sense vector model remarkably well.

1 Introduction

Distributional word representations based on counting co-occurrences have a long history in natural language processing and have successfully been applied to numerous tasks such as sentiment analysis, recognising textual entailment, word-sense disambiguation and many other important problems. More recently low-dimensional and dense neural word embeddings have received a considerable amount of attention in the research community and have become ubiquitous in numerous NLP pipelines in academia and industry. One fundamental simplifying assumption commonly made in distributional semantic models, however, is that every word can be encoded by a single representation. Combining polysemous lexemes into a single vector has the consequence of essentially creating a weighted average of all observed meanings of a lexeme in a given text corpus.

Therefore a number of proposals have been made to overcome the issue of conflating several different senses of an individual lexeme into a single representation. One approach (Reisinger and Mooney, 2010; Huang et al., 2012) is to try directly inferring a predefined number of senses from data and subsequently label any occurrences of a polysemous lexeme with the inferred inventory. Similar approaches are proposed by Reddy et al. (2011) and Kartsaklis et al. (2013) who show that appropriate sense selection or disambiguation typically improves performance for composition of noun phrases (Reddy et al., 2011) and verb phrases (Kartsaklis et al., 2013). Dinu and Lapata (2010) proposed a model that represents the meaning of a word as a probability distribution over latent senses which is modulated based on contextualisation and report improved performance on a word similarity task and the lexical substitution task. Other approaches leverage an existing lexical resource such as BabelNet or WordNet to obtain sense labels *a priori* to creating word representations (Iacobacci et al., 2015), or as a postprocessing step after obtaining initial word representations (Chen et al., 2014; Pilehvar and Collier, 2016). While these approaches have exhibited strong performance on benchmark word similarity tasks (Huang et al., 2012; Iacobacci et al., 2015) and some downstream processing tasks such as part-of-speech tagging and relation identification (Li and Jurafsky, 2015), they have been weaker than the single-vector representations at predicting the compositionality of multi-word expressions (Salehi et al., 2015), and at tasks which require the meaning of a word to be considered in context; e.g, word sense disambiguation (Iacobacci et al., 2016) and word similarity in context (Iacobacci et al., 2015).

In this paper we consider what happens when distributional representations are composed to

form representations for larger units of meaning. In a compositional phrase, the meaning of the whole can be inferred from the meaning of its parts. Thus, assuming compositionality, the representation of a phrase such as *black mood*, should be directly inferable from the representations for *black* and for *mood*. Further, one might suppose that composing the correct senses of the individual lexemes would result in a more accurate representation of that phrase. However, our counter-hypothesis is that the act of composition contextualises or disambiguates each of the lexemes thereby making the representations of individual senses redundant. We investigate this hypothesis by evaluating the performance of single-vector representations and multi-sense representations at both a benchmark phrase similarity task and at a novel word-sense discrimination task.

Our contributions in this work are thus as follows. First, we provide quantitative and qualitative evidence that even simple composition functions have the ability to recover sense-specific information from a single-vector representation of a polysemous lexeme in context. Second, we introduce a novel word-sense discrimination task[1], which can be seen as the first stage of word-sense disambiguation. The goal is to find whether the occurrences of a lexeme in two or more sentential contexts belong to the same sense or not, without necessarily labelling the senses. While it has received relatively little attention in recent years, it is an important natural language understanding problem and can provide important insights into the process of semantic composition.

2 Evaluating Distributional Models of Composition

For evaluation we use several readily available off-the-shelf word embeddings, that have already been shown to work well for a number of different NLP applications. We compare the 300-dimensional skip-gram `word2vec` (Mikolov et al., 2013) word embeddings[2] to the dependency based version of `word2vec` — henceforth `dep2vec`[3] (Levy and Goldberg, 2014) — and the SENSEMBED model[4] by Iacobacci et al. (2015), which creates word-sense embeddings by performing word-sense disambiguation prior to running `word2vec`.

We note that `word2vec` and `dep2vec` use a single vector per word approach and therefore conflate the different senses of a polysemous lexeme. On the other hand, SENSEMBED utilises Babelfy (Moro et al., 2014) as an external knowledge source to perform word-sense disambiguation and subsequently creates one vector representation per word sense.

For composition we use pointwise addition for all models as this has been shown to be a strong baseline in a number of studies (Hashimoto et al., 2014; Hill et al., 2016). We also experimented with pointwise multiplication as composition function but, similar to Hill et al. (2016), found its performance to be very poor (results not reported). We model any out-of-vocabulary items as a vector consisting of all zeros and determine proximity of composed meaning representations in terms of cosine similarity. We lowercase and lemmatise the data in our task but do not perform number or date normalisation, or removal of rare words.

3 Phrase Similarity

Our first evaluation task is the benchmark phrase similarity task of Mitchell and Lapata (2010). This dataset consists of 108 adjective-noun (AN), 108 noun-noun (NN) and 108 verb-object (VO) pairs. The task is to compare a compositional model's similarity estimates with human judgements by computing Spearman's ρ. An average ρ of 0.47-0.48 represents the current state-of-the-art performance on this task (Hashimoto et al., 2014; Kober et al., 2016; Wieting et al., 2015).

For single-sense representations, the strategy for carrying out this task is simple. For each phrase in each pair, we compose the constituent representations and then compute the similarity of each pair of phrases using the cosine similarity. For multi-sense representations, we adapted the strategy which has been used successfully in various word similarity experiments (Huang et al., 2012; Iacobacci et al., 2015). Typically, for each word pair, all pairs of senses are considered and the similarity of the word pair is considered to be

[1] Our task is available from `https://github.com/tttthomasssss/sense2017`

[2] Available from: `https://code.google.com/p/word2vec/`

[3] Available from: `https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/`

[4] Available from: `http://lcl.uniroma1.it/sensembed/`

the similarity of the closest pair of senses. The fact that this strategy works well suggests that when humans are asked to judge word similarity, the pairing automatically primes them to select the closest senses. Extending this to phrase similarity requires us to compose each possible pair of senses for each phrase and then select the sense configuration which results in maximal phrase similarity. For comparison, we also give results for the configuration which results in minimal phrase similarity and the arithmetic mean[5] of all sense configurations.

3.1 Results

Model	AN	NN	VO	Average
word2vec	0.47	**0.46**	**0.45**	**0.46**
dep2vec	**0.48**	**0.46**	**0.45**	**0.46**
SENSEMBED:max	0.39	0.39	0.32	0.37
SENSEMBED:min	0.24	0.12	0.22	0.19
SENSEMBED:mean	0.46	0.35	0.37	0.39

Table 1
Results for the Mitchell and Laptata (2010) dataset.

Table 1 shows that the simple strategy of adding high quality single-vector representations is very competitive with the state-of-the-art for this task (Hashimoto et al., 2014). None of the strategies for selecting a sense configuration for the multi-sense representations could compete with the single sense representations on this task. One possible explanation is that the commonly adopted closest sense strategy is not effective for composition since the composition of incorrect senses may lead to spuriously high similarities (for two "implausible" sense configurations).

Table 2 lists a number of example phrase pairs with low average human similarity scores in the Mitchell and Lapata (2010) test set. The results show the tendency of the closest sense strategy with SENSEMBED (SE) to overestimate the similarity of dissimilar phrase pairs. For a comparison we manually labelled the lexemes in the sample phrases with the appropriate BabelNet senses prior to composition (SE*). Human (H) similarity scores are normalised and averaged for an easier comparison, model estimates represent cosine similarities.

[5]We also tried the geometric mean and the median but these performed comparably with the arithmetic mean.

Phrase 1	Phrase 2	SE	SE*	H
buy land	*leave house*	0.49	0.28	0.26
close eye	*stretch arm*	0.40	0.31	0.25
wave hand	*leave company*	0.42	0.08	0.20
drink water	*use test*	0.29	0.04	0.18
european state	*present position*	0.28	-0.03	0.19
high point	*particular case*	0.41	0.10	0.21

Table 2
Tendency of SENSEMBED (SE) to overestimate the similarity on phrase pairs with low average human similarity when the closest sense strategy is used.

4 Word Sense Discrimination

Word-sense discrimination can be seen as the first stage of word-sense disambiguation, where the goal is to find whether two or more occurrences of the same lexeme express identical senses, without necessarily labelling the senses yet. It has received relatively little attention despite its potential for providing important insights into semantic composition, focusing in particular on to the ability of compositional distributional semantic models to appropriately contextualise a polysemous lexeme.

Work on word-sense discrimination has suffered from the absence of a benchmark task as well as a clear evaluation methodology. For example Schütze (1998) evaluated his model on a dataset consisting of 20 polysemous words (10 naturally ambiguous lexemes and 10 artificially ambiguous "pseudo-lexemes") in terms of accuracy for coarse grained sense distinctions, and an information retrieval task. Pantel and Lin (2002), and Van de Cruys (2008) used automatically extracted words from various newswire sources and evaluated the output of their models in comparison to WordNet and EuroWordNet, respectively. Purandare and Pedersen (2004) used a subset of the words from the SENSEVAL-2 task and evaluated their models in terms of precision, recall and F1-score of how well available sense tags match with clusters discovered by their algorithms. Akkaya et al. (2012) used the concatenation of the SENSEVAL-2 and SENSEVAL-3 tasks and evaluated their models in terms of cluster purity and accuracy. Finally, Moen et al. (2013) use the semantic textual similarity (STS) 2012 task, which is based on human judgements of the similarity between two sentences.

One contribution of our work is a novel word-sense discrimination task, evaluated on a number of robust baselines in order to facilitate future research in that area. In particular, our task offers a testbed for assessing the contextualisation ability

of compositional distributional semantic models. The goal is, for a given polysemous lexeme in context, to identify the sentence from a list of options that is expressing the same sense of that lexeme as the given target sentence. These two sentences — the target and the "correct answer" — can exhibit any degree of semantic similarity as long as they convey the same sense of the target lexeme. Table 3 shows an example of the polysemous adjective *black* in our task. The goal of any model would be to determine that the expressed sense of *black* in the sentence *She was going to set him free from all of the evil and black hatred he had brought to the world* is identical to the expressed sense of *black* in the target sentence *Or should they rebut the Democrats' black smear campaign with the evidence at hand*.

Our task assesses the ability of a model to discriminate a particular sense in a sentential context from any other senses and thus provides an excellent testbed for evaluating multi-sense word vector models as well as compositional distributional semantic models. By composing the representation of a target lexeme with its surrounding context, it should be possible to determine it's sense. For example, composing *black smear campaign* should lead to a compositional representation that is closer to the composed representation of *black hatred* than to *black mood*, *black sense of humour* or *black coffee*. This essentially uses the similarity of the compositional representation of a lexeme's context to determine its sense. Similar approaches to word-sense disambiguation have already been successfully used in past works (Akkaya et al., 2012; Basile et al., 2014).

4.1 Task Construction

For the construction of our dataset we made use of data from two english dictionaries (Oxford Dictionary and Collins Dictionary), accessible via their respective web APIs[6], as well as examples from the sense annotated corpus SemCor (Miller et al., 1993). Our use of dictionary data is motivated by a number of favourable properties which make it a very suitable data source for our proposed task:

- The content is of very high-quality and curated by expert lexicographers.
- All example sentences are carefully crafted in order to unambiguously illustrate the usage of a particular sense for a given polysemous lexeme.
- The granularity of the sense inventory reflects common language use[7].
- The example sentences are typically free of any domain bias wherever possible.
- The data is easily accessible via a web API.

By using the data from curated resources we were able to avoid a setup as a sentence similarity task and any potentially noisy crowd-sourced human similarity judgements.

We were furthermore able to collect data from varying frequency bands, enabling an assessment of the impact of frequency on any model. Figure 1 shows the number of target lexemes per frequency band. While the majority of lexemes, with reference to a cleaned October 2013 Wikipedia dump[8], is in the middle band, there is a considerable amount of less frequent lexemes. The most frequent target lexeme in our task is the verb *be* with $\approx$1.8m occurrences in Wikipedia, and the least frequent lexeme is the verb *ruffle* with only 57 occurrences. The average target lexeme frequency is $\approx$95k for adjectives, and $\approx$45k$-$46k for nouns and verbs[9].

4.2 Task Setup Details

We collected data for 3 different parts-of-speech: adjectives, nouns and verbs. We furthermore created task setups with varying numbers of senses to distinguish (2-5 senses) for a given target lexeme. This is to evaluate how well a model is able to discriminate different degrees of polysemy of any lexeme. For any task setup evaluating for n senses, we included all lexemes with $> n$ senses and randomly sampled n senses from its inventory. For each lexeme, we furthermore ensured that it had at least 2 example sentences per sense. For the available senses of any given lexeme, we randomly chose a sense as the target sense, and from

[6] `https://developer.oxforddictionaries.com` for the Oxford Dictionary, `https://www.collinsdictionary.com/api/` for the Collins Dictionary. We use NLTK 3.2 to access SemCor.

[7] The Oxford dictionary lists 5 different senses for the noun "bank", whereas WordNet 3.0 lists 10 synsets, for example distinguishing "bank" as the concept for a financial institution and "bank" as a reference to the building where financial transactions take place.

[8] We removed any articles with fewer than 20 page views.

[9] The overall number of unique word types is smaller than the number of examples in our task as there are a number of lexemes that can occur with more than one part-of-speech.

	Sense Definition	Sentence
Target	full of anger or hatred	Or should they rebut the Democrats' **black** smear campaign with the evidence at hand?
Option 1	full of anger or hatred	She was going to set him free from all of the evil and **black** hatred he had brought to the world.
Option 2	(of a person's state of mind) full of gloom or misery; very depressed	I've been in a **black** mood since September 2001, it's hanging over me like a penumbra.
Option 3	(of humour) presenting tragic or harrowing situations in comic terms	Over the years I have come to believe that fate either hates me, or has one hell of a **black** sense of humour.
Option 4	(of coffee or tea) served without milk	The young man was reading a paperback novel and sipping a steaming mug of hot, **black** coffee.

Table 3: Example of the polysemous adjective *black* in our task. The goal for any model is to predict option 1 as expressing the same sense of *black* as the target sentence.

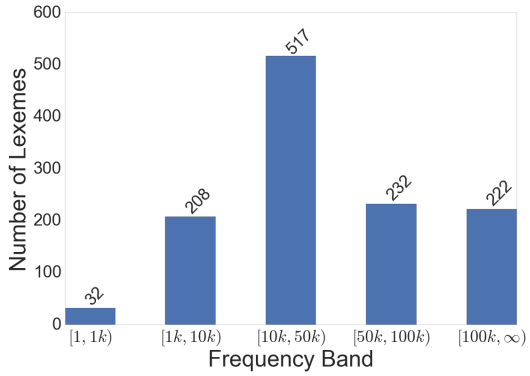

Figure 1: Binned frequency distribution of the polysemous target lexemes in our task.

	2 senses	3 senses	4 senses	5 senses
Adjective	66/209	47/170	37/137	28/115
Noun	170/618	125/499	100/412	74/345
Verb	127/438	71/354	72/295	56/256
Total	363/1265	263/1023	209/844	164/716

Table 4: Number of examples per part-of-speech and number of senses (*#dev examples*/*#train examples*).

its list of example sentences randomly sampled 2 sentences, one as the target example and one as the "correct answer" for the list of candidate sentences. Finally we once again randomly sampled the required number of other senses and example sentences to complete the task setup. Using random sampling of word senses and targets aims to avoid a predominant sense bias.

For each part-of-speech we created a development split for parameter tuning and a test split for the final evaluation. Table 4 shows the number of examples for each setup variant of our task. The two biggest categories, making up the majority of the dataset, are polysemous nouns, representing roughly half of the data, and the smallest category are adjectives. It is worth noting that each example consists of $n+1$ sentences, where n is the number of senses. That means that the total number of sentences for the 3 sense setup, for example, is 1052 sentences for development and 4092 sentences for testing. We measure performance in terms of accuracy of correctly predicting which two sentences share the same sense of a given target lexeme. Accuracy has the advantage of being much easier to interpret — in absolute terms as well as in the relative difference between models — in comparison to other commonly used evaluation metrics such as cluster purity measures or correlation metrics such as Spearman ρ and Pearson r.

4.3 Experimental Setup

In this paper we compare the compositional models outlined earlier with two baselines, a random baseline and a word-overlap baseline of the extracted contexts. For the single-vector representations, we composed the target lexeme with all of the words in the context window and compared it with the equivalent representation of each of the options (lexeme plus context words). The option with the highest cosine similarity was deemed to be the selected sense. For SENSEMBED, we composed all sense vectors of a target lexeme with the given context and then used the *closest sense* strategy (Iacobacci et al., 2015) on composed representations to choose the predicted sense[10]. The word-overlap baseline is simply the number of words in common between the context window for the target and each of the options.

We experimented with symmetric linear bag-of-words contexts of size 1, 2 and 4 around the target lexeme. We also experimented with dependency contexts, where first-order dependency contexts performed almost identical to using a 2-word bag-of-words context window (results not

[10]We also tried an all-by-all senses composition, however found this to be computationally not tractable.

reported). We excluded stop words prior to extracting the context window in order to maximise the number of content words. We break ties for any of the methods — including the baselines — by randomly picking one of the options with the highest similarity to the composed representation of the target lexeme with its context. Statistical significance between the best performing model and the word overlap baseline is computed by using a randomised pairwise permutation test (Efron and Tibshirani, 1994).

4.4 Results

Table 5 shows the results for all context window sizes across all parts-of-speech and number of senses. All models substantially outperform the random baseline for any number of senses. Interestingly the word overlap baseline is competitive for all context window sizes. While it is a very simple method, it has already been found to be a strong baseline for paraphrase detection and semantic textual similarity (Dinu and Thater, 2012). One possible explanation for its robust performance on our task is an occurrence of the one-sense-per-collocation hypothesis (Yarowsky, 1993). The performance of all other models is roughly in the same ballpark for all parts-of-speech and number of senses, suggesting that they form robust baselines for future models. While the results are relatively mixed for adjectives, `word2vec` appears to be the strongest model for polysemous nouns and verbs.

The perhaps most interesting observation in Table 5 is that `word2vec` and `dep2vec` are performing as well or even better than SENSEMBED despite the fact that the former conflate the senses of a polysemous lexeme in a single vector representation. Figure 2 shows the average performance of all models across parts-of-speech per number of senses and for all context window sizes. All models exhibit a strong performance across the board, making them robust baselines for future work on this task.

SENSEMBED and Babelfy

One possible explanation for SENSEMBED not outperforming the other methods despite its cleaner encoding of different word senses in the above experiments is that at train time, it had access to sense labels from Babelfy. At test time on our task however, it did not have any sense labels available. We therefore sense tagged the 5-sense noun subtask with Babelfy and re-ran SENSEMBED. As Table 6 shows, access to sense labels at test time did not give a substantive performance boost, representing further support for our hypothesis that composition in single-sense vector models might be sufficient to recover sense specific information.

Frequency Range

We chose the 2-sense noun subtask to estimate the degree sensitivity of target lexeme frequency on our task we merged the $[1, 1k)$ and $[1k, 10k)$, and $[50k, 100k)$ and $[100k, \infty)$ frequency bands from Figure 1, and sampled an equal number of target words from each band. Table 7 reports the results for this experiment. All methods outperform the random and word overlap baseline and appear to be working better for less frequent lexemes. One possible explanation for this behaviour is that less frequent lexemes have fewer senses and potentially less subtle sense differences than more frequent lexemes, which would make them easier to discriminate by distributional semantic methods.

5 Discussion

Our results suggest that pointwise addition in a single-sense vector model such as `word2vec` is able to discriminate the sense of a polysemous lexeme in context in a surprisingly effective way and represents a strong baseline for future work. Distributional composition can therefore be interpreted as a process of *contextualising* the meaning of a lexeme. This way, composition does not only act as a way to represent the meaning of a phrase as a whole, but also as a local discriminator for any lexemes in the phrase. For example the composed representation of *dry clothes* should only keep contexts that *dry* shares with *clothes* while suppressing contexts it shares with *weather* or *wine*. Hence, one would expect the same to happen with a polysemous lexeme such as *bank* in the context of *river* and *account*, respectively.

Recent work by Arora et al. (2016) has shown that the different senses of a polysemous lexeme reside in a linear substructure within a single vector and are recoverable by sparse coding, which they relate to as a soft clustering equivalent in linear algebra. There is furthermore evidence that additive composition in low-dimensional word embeddings approximates an intersection of the contexts of two distributional word vectors (Tian et

Symmetric context window of size 1

	Adjective				Noun				Verb			
Senses	2	3	4	5	2	3	4	5	2	3	4	5
Random	0.53	0.32	0.25	0.14	0.47	0.32	0.23	0.19	0.47	0.31	0.23	0.18
Word Overlap	0.63	0.46	0.47	0.40	0.55	0.40	0.37	0.34	0.54	0.44	0.38	0.29
word2vec	**0.70**	0.56	**0.61**†	0.54†	**0.66**‡	**0.52**‡	**0.50**‡	0.44‡	**0.63**‡	**0.56**‡	**0.52**‡	**0.43**‡
dep2vec	0.65	**0.64**‡	0.57	**0.57**‡	0.64‡	0.50‡	0.49‡	**0.48**‡	**0.63**‡	0.55‡	0.50‡	**0.43**‡
SENSEMBED	0.67	0.54	0.56	0.56†	0.64‡	0.49‡	**0.50**‡	0.43‡	0.62†	0.53‡	0.49‡	0.38†

Symmetric context window of size 2

	Adjective				Noun				Verb			
Senses	2	3	4	5	2	3	4	5	2	3	4	5
Random	0.53	0.32	0.25	0.14	0.47	0.32	0.23	0.19	0.47	0.31	0.23	0.18
Word Overlap	0.66	0.51	0.55	0.43	0.59	0.47	0.43	0.41	0.58	0.51	0.45	0.36
word2vec	0.70	0.64†	0.58	0.55	**0.71**‡	**0.63**‡	**0.59**‡	0.54‡	**0.68**‡	0.64‡	**0.58**‡	**0.49**‡
dep2vec	0.71	**0.65**‡	0.58	**0.57**‡	0.70‡	0.57‡	0.55‡	**0.55**‡	0.66†	0.64‡	0.54†	0.46†
SENSEMBED	**0.72**‡	0.62	**0.61**	0.52	0.69‡	0.56†	0.57‡	0.51†	0.67‡	**0.65**†	0.57	0.45

Symmetric context window of size 4

	Adjective				Noun				Verb			
Senses	2	3	4	5	2	3	4	5	2	3	4	5
Random	0.53	0.32	0.25	0.14	0.47	0.32	0.23	0.19	0.47	0.31	0.23	0.18
Word Overlap	0.67	0.55	0.58	0.51	0.62	0.50	0.49	0.45	0.59	0.55	0.50	0.40
word2vec	0.71	0.65†	**0.65**	**0.57**	**0.73**‡	**0.61**‡	**0.62**‡	**0.57**‡	**0.71**‡	**0.62**†	**0.57**	**0.53**‡
dep2vec	0.72	**0.66**†	0.60	0.54	0.71‡	0.55	0.56†	0.53†	0.67	0.62	0.54	0.50
SENSEMBED	**0.75**	0.59	0.62	0.55	0.69‡	0.57†	0.58‡	0.53†	0.68‡	0.62†	0.55	0.47

Table 5
Performance overview for all parts-of-speech and number of senses, ‡ statistically significant at the $p < 0.01$ level in comparison to the Word Overlap baseline; † statistically significant at the $p < 0.05$ level in comparison to the Word Overlap baseline.

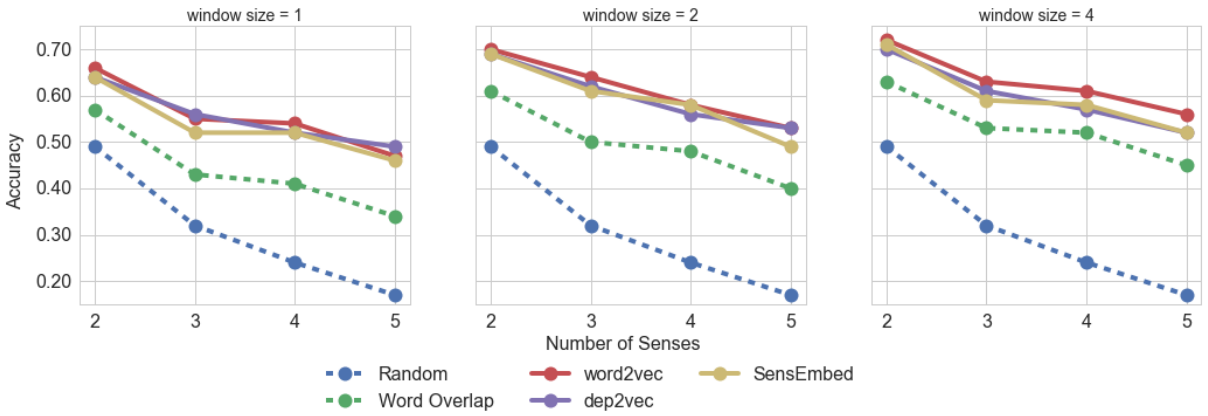

Figure 2: Average performance across parts-of-speech per number of senses and context window.

Noun - 5 Senses

Context Window Size	**1**	**2**	**4**
word2vec	0.44	0.54	**0.57**
dep2vec	**0.48**	**0.55**	0.53
SENSEMBED	0.43	0.51	0.53
SENSEMBED & Babelfy	0.45	0.49	0.54

Table 6
Results on the 5-sense noun subtask with SENSEMBED having access to Babelfy sense labels at test time.

Noun - 2 Senses, context window size = 2

Frequency Band	$< 10k$	$10k - 50k$	$\geq 50k$
Random	0.51	0.51	0.51
Word Overlap	0.66	0.60	0.56
word2vec	**0.81**	0.64	**0.66**
dep2vec	0.77	0.67	**0.66**
SENSEMBED	0.74	**0.68**	0.60

Table 7
Results on a subsample of the 2-sense noun subtask across frequency bands.

al., 2015). It therefore seems plausible that an intersective composition function should be able to recover sense specific information.

To qualitatively analyse this hypothesis we used the word2vec and SENSEMBED vectors to compose a small number of example phrases

by pointwise addition and calculated their top 5 nearest neighbours in terms of cosine similarity. For SENSEMBED we manually sense tagged the phrases with the appropriate BabelNet sense labels prior to composition. Note that we omitted the BabelNet sense labels in the neighbour list for brevity, however they were consistent with the intended sense in all cases. Table 8 supports the view of composition as a way of contextualising the meaning of a lexeme. In all cases in our example the `word2vec` neighbours reflect the intended sense of the polysemous lexeme, providing evidence for the linear substructure of word senses in a single vector as discovered by Arora et al. (2016), and suggesting that distributional composition is able to recover sense specific information from a polysemous lexeme. Furthermore it seems natural that the very fine-grained sense-level vector space of SENSEMBED is giving rise to a very focused neighbourhood, however there does not seem to be a clear advantage over `word2vec` from a qualitative point of view when using simple additive composition.

6 Related Work

The perhaps most popular tasks for evaluating the ability of a model to capture or encode the different senses of a polysemous lexeme in a given context are the english lexical substitution task (McCarthy and Navigli, 2007) and the Microsoft sentence completion challenge (Zweig and Burges, 2011). Both tasks require any model to fill an appropriate word into a pre-defined slot in a given sentential context. The sentence completion challenge provides a list of candidate words while the english lexical substitution task does not. However, neither task focuses on polysemy and the english lexical substitution task conflates the problems of discriminating word senses and finding meaning preserving substitutes.

Dictionary definitions have previously been used to evaluate compositional distributional semantic models where the goal is to match a dictionary entry with its corresponding definition (Kartsaklis et al., 2012; Polajnar and Clark, 2014). These datasets are commonly set up as retrieval tasks, but generally do not test the ability of a model to disambiguate a polysemous word in context, or discriminate multiple definitions of the same word.

Our task also provides a novel evaluation for compositional distributional semantic models, where the predominant strategy is to estimate the similarity of two short phrases (Mitchell and Lapata, 2008; Mitchell and Lapata, 2010; Grefenstette and Sadrzadeh, 2011; Turney, 2012; Bernardi et al., 2013; Kartsaklis and Sadrzadeh, 2014) or sentences (Huang et al., 2012; Pham et al., 2013; Marelli et al., 2014; Agirre et al., 2016) in comparison to human provided gold-standard judgements. One problem with these similarity tasks is that the similarity or relatedness of two sentences is very difficult to judge — especially on a fine-grained scale — even for humans. This frequently results in a relatively high variance of judgements and low inter-annotator agreement (Batchkarov et al., 2016). The short phrase datasets typically have a fixed structure that only test a very small fraction of the possible grammatical constructions in which a lexeme can occur, and furthermore provide very little context. The use of full sentences remedies the lack of context and grammatical variation, however can still contain a significant level of noise due to the automatic construction of the dataset or the variance in human ratings. In contrast, our task is not set up as a sentence similarity task and therefore avoids the use of human similarity judgements.

Our task is similar to word-sense induction (WSI), however we only focus on discriminating the sense of a polysemous lexeme in context rather than inducing a set of senses from raw data and appropriately tagging subsequent occurrences of polysemous instances with the inferred inventory. Separating the sense discrimination task from the problem of sense induction has the advantage of making our task applicable to evaluating compositional distributional semantic models in order to test their ability to appropriately contextualise a polysemous lexeme. Due to not requiring any models to perform an extra step for sense induction, our task is easier to evaluate as no matching between sense clusters identified by a model and some gold standard sense classes needs to be performed, as typically proposed in the WSI literature (Agirre and Soroa, 2007; Manandhar et al., 2010).

Most closely related to our task are the Stanford Contextual Word Similarity (SCWS) dataset by Huang et al. (2012) and the Usage Similarity (USim) task by Erk et al. (2009). The goal in both tasks is to estimate the similarity of two

Phrase	word2vec neighbours	SENSEMBED neighbours
river bank	bank, river, creek, lake, rivers	bank, river, stream, creek, river basin
bank account	account, bank, accounts, banks, citibank	bank, banks, the bank, pko bank polski, handlowy
dry weather	weather, dry, wet weather, wet, unreasonably warm	dry, weather, humid, cold, cool
dry clothes	dry, clothes, clothing, rinse thoroughly, wet	dry, clothes, warm, cold, wet
capital city	capital, city, cities, downtown, town	city, capital, the capital city, town, provincial capital
capital asset	capital, asset, assets, investment, worth	capital, asset, investment, assets, investor
power plant	plant, power, plants, coalfired, megawatt	power, plant, near-limitless, pulse-power, power of the wind
garden plant	plant, garden, plants, gardens, vegetable garden	plant, garden, plants, oakville assembly, solanaceous
window bar	bar, window, windows, doorway, door	window, bar, windows, glass window, wall
sandwich bar	bar, sandwich, restaurant, burger, diner	sandwich, bar, restaurant, hot dog, cake
gasoline tank	gasoline, tank, fuel, gallon, tanks	gasoline, tank, fuel, petrol, kerosene
armored tank	armored, tank, tanks, M1A1 Abrams, armored vehicle	armored, armoured, tank, tanks, light tank
desert rock	rock, desert, rocks, desolate expanse, arid desert	desert rock, the desert, deserts, badlands
rock band	rock, band, rockers, bands, indie rock	band, rock, group, the band, rock group

Table 8
Nearest neighbours of composed phrases for word2vec and SENSEMBED. Distributional composition in word2vec is able to recover sense specific information remarkably well. Some neighbours are phrases because they have been encoded as a single token in the original vector space.

polysemous words in context in comparison to human provided gold standard judgements. In the SCWS dataset typically two different lexemes are considered whereas in USim and our task the same lexemes with different contexts are compared. Instead of relying on crowd-sourced human gold-standard similarity judgements, which can be prone to a considerable amount of noise[11], we leverage the high-quality content of available english dictionaries. Furthermore, our task is not formulated as estimating the similarity between two lexemes in context, but identifying the sentences that use the same sense of a given polysemous lexeme.

7 Conclusion

While elementary multi-sense representations of words might capture a more fine grained semantic picture of a polysemous word, that advantage does not appear to transfer to distributional composition in a straightforward way. Our experiments on a popular phrase similarity benchmark and our novel word-sense discrimination task have demonstrated that semantic composition does not appear to benefit from a fine grained sense inventory, but that the ability to contextualise a polysemous lexeme in single-sense vector models is sufficient for superior performance. We furthermore have provided qualitative and quantitative evidence that an intersective composition function such as pointwise addition for neural word embeddings is able to discriminate the meaning of a word in context, and is able to recover sense specific information remarkably well.

Lastly, our experiments have uncovered an important question for multi-sense vector models, namely how to exploit the fine-grained sense level representations for distributional composition. Our novel word-sense discrimination task provides an excellent testbed for compositional distributional semantic models, both following a single-sense or multi-sense vector modelling paradigm, due to its focus on assessing the ability of a model to appropriately contextualise the meaning of a word. Our task furthermore provides another evaluation option away from intrinsic evaluations which are based on often noisy human similarity judgements, while also not being embedded in a downstream task.

In future work we aim to extend our evaluation to more complex compositional distributional semantic models such as the lexical function model (Paperno et al., 2014) or the Anchored Packed Dependency Tree framework (Weir et al., 2016). We furthermore want to investigate how far the sense-discriminating ability of composition can be leveraged for other tasks.

[11] For example the average standard deviation of human ratings in the SCWS dataset is ≈3 on a 10-point scale, and can be up to 4–5 in some cases.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic, June. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-

2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California, June. Association for Computational Linguistics.

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. 2012. Utilizing semantic composition in distributional semantic models for word sense discrimination and word sense disambiguation. In *Proceedings of ICSC*, pages 45–51.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. Linear algebraic structure of word senses, with applications to polysemy. *CoRR*, abs/1601.03764.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1591–1600, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method of evaluating distributional semantic models. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 7–12. Association for Computational Linguistics.

Raffaella Bernardi, Georgiana Dinu, Marco Marelli, and Marco Baroni. 2013. A relatedness benchmark to test the role of determiners in compositional distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 53–57, Sofia, Bulgaria, August. Association for Computational Linguistics.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1025–1035.

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

Georgiana Dinu and Stefan Thater. 2012. Saarland: Vector-based models of semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 603–607, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Bradley Efron and Robert Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC press.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18, Suntec, Singapore, August. Association for Computational Linguistics.

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

Kazuma Hashimoto, Pontus Stenetorp, Makoto Miwa, and Yoshimasa Tsuruoka. 2014. Jointly learning word representations and composition functions using predicate-argument structures. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1544–1555, Doha, Qatar, October. Association for Computational Linguistics.

Felix Hill, KyungHyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea, July. Association for Computational Linguistics.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105, Beijing, China, July. Association for Computational Linguistics.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany, August. Association for Computational Linguistics.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2014. A study of entanglement in a categorical framework of

natural language. In *Proceedings of the 11th Workshop on Quantum Physics and Logic (QPL)*.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2012. A unified sentence space for categorical distributional-compositional semantics: Theory and experiments. In *Proceedings of COLING 2012: Posters*, pages 549–558, Mumbai, India, December. The COLING 2012 Organizing Committee.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2013. Separating disambiguation from composition in distributional semantics. In *Proceedings of CoNLL*, pages 114–123. Association for Computational Linguistics.

Thomas Kober, Julie Weeds, Jeremy Reffin, and David Weir. 2016. Improving sparse word representations with distributional inference for semantic composition. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1691–1702, Austin, Texas, November. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland, June. Association for Computational Linguistics.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732, Lisbon, Portugal, September. Association for Computational Linguistics.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. Semeval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden, July. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1314.

Diana McCarthy and Robert Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the Arpa Workshop on Human Language Technology*, pages 303–308.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technology Conference*, pages 236–244, Columbus, Ohio, June. Association for Computational Linguistics.

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

Hans Moen, Erwin Marsi, and Björn Gambäck. 2013. Towards dynamic word sense discrimination with random indexing. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 83–90, Sofia, Bulgaria, August. Association for Computational Linguistics.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 613–619, New York, NY, USA. ACM.

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. Association for Computational Linguistics.

Nghia Pham, Raffaella Bernardi, Yao Zhong Zhang, and Marco Baroni. 2013. Sentence paraphrase detection: When determiners and word order make the difference. In *Proceedings of IWCS 2013 Workshop Towards a Formal Distributional Semantics*, pages 21–29, Potsdam, Germany, March. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Nigel Collier. 2016. De-conflated semantic representations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1680–1690, Austin, Texas, November. Association for Computational Linguistics.

Tamara Polajnar and Stephen Clark. 2014. Improving distributional semantic vectors through context selection and normalisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 230–238, Gothenburg, Sweden, April. Association for Computational Linguistics.

Amruta Purandare and Ted Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of CoNLL*, pages 41–48. Association for Computational Linguistics.

Siva Reddy, Ioannis Klapaftis, Diana McCarthy, and Suresh Manandhar. 2011. Dynamic and static prototype vectors for semantic composition. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 705–713, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Proceedings of NAACL-HLT*, pages 109–117. Association for Computational Linguistics.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado, May–June. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, mar.

Ran Tian, Naoaki Okazaki, and Kentaro Inui. 2015. The mechanism of additive composition. *CoRR*, abs/1511.08407.

Peter D. Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44(1):533–585, may.

Tim Van de Cruys. 2008. Using three way data for word sense discrimination. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 929–936, Manchester, UK, August. Coling 2008 Organizing Committee.

David Weir, Julie Weeds, Jeremy Reffin, and Thomas Kober. 2016. Aligning packed dependency trees: a theory of composition for distributional semantics. *Computational Linguistics, special issue on Formal Distributional Semantics*, 42(4):727–761, December.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *TACL*, 3:345–358.

David Yarowsky. 1993. One sense per collocation. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, pages 266–271, Stroudsburg, PA, USA. Association for Computational Linguistics.

Geoffrey Zweig and J.C. Chris Burges. 2011. The microsoft research sentence completion challenge. Technical report, Microsoft Research.