Coming to Your Senses: on Controls and Evaluation Sets in Polysemy Research

Haim Dubossarsky[1], **Eitan Grossman**[2] and **Daphna Weinshall**[3]
[1] Edmond and Lily Safra Center for Brain Sciences
[2] Department of Linguistics
[3] School of Computer Science and Engineering
The Hebrew University of Jerusalem, 91904 Jerusalem, Israel
haim.dub@gmail.com, {eitan.grossman,daphna}@mail.huji.ac.il

Abstract

The point of departure of this article is the claim that sense-specific vectors provide an advantage over normal vectors due to the polysemy that they presumably represent. This claim is based on performance gains observed in gold standard evaluation tests such as word similarity tasks. We demonstrate that this claim, at least as it is instantiated in prior art, is unfounded in two ways. Furthermore, we provide empirical data and an analytic discussion that may account for the previously reported improved performance. First, we show that ground-truth polysemy degrades performance in word similarity tasks. Therefore word similarity tasks are not suitable as an evaluation test for polysemy representation. Second, random assignment of words to senses is shown to improve performance in the same task. This and additional results point to the conclusion that performance gains as reported in previous work may be an artifact of random sense assignment, which is equivalent to sub-sampling and multiple estimation of word vector representations. Theoretical analysis shows that this may on its own be beneficial for the estimation of word similarity, by reducing the bias in the estimation of the cosine distance.

1 Introduction

Polysemy is a fundamental feature of natural languages, which typically have many polysemic words. *Chair*, for example, can refer to either a piece of furniture or to a person in charge of a meeting. Therefore both theoretical linguistics and computational linguistics seek to establish principled methods of identifying the senses that together constitute the meaning of words.

It is commonly assumed or claimed that standard word embeddings are unable to capture polysemy (Iacobacci et al., 2015), which results in suboptimal performance in gold standard evaluation tests such as word similarity tasks, and potentially hamper performance in downstream tasks. The corollary assumption is that sense-specific representations will lead to improved performance on these evaluation tests. This assumption is conceptually attractive, since it makes sense that sense-specific representations are more accurate than global representations which conflate the different senses of a word into a single representation. For example, in translating 'chair,' it is reasonable that performance should improve if the two senses are represented separately. This view is supported by several studies (Huang et al., 2012; Neelakantan et al., 2014; Chen et al., 2014; Li and Jurafsky, 2015; Iacobacci et al., 2015; Mancini et al., 2017), which argue that sense-specific representations lead to improved performance in word similarity tasks.

Ideally, such claims about polysemy should be evaluated using a gold standard evaluation set that is tailored specifically for polysemous words. As this set does not exist, tasks involving word similarity tests have been used as a proxy (see Section 2). The underlying hypothesis is that enriching word vector representations with polysemic information should express itself in performance gains in these tasks. However, this hypothesis has never been tested directly, and the ability of word similarity tasks to directly benefit from polysemic information must first be validated if they are to serve as genuine evaluation sets in polysemy research. Until then, the validity of any reported positive effects of sense-specific representations on evaluation tests is to be treated with caution.

In this paper our first aim is to assess the validity of *word similarity tasks* as proper evaluation tests for polysemic word representations. We use two independent corpora in order to obtain polysemic vectors: (i) a sense-annotated corpus, and (ii) an artificially-induced annotated corpus, constructed

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1732–1740
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

by using an established method that we modify for our purposes. Surprisingly, our analyses show that even the most accurate sense-specific word vectors do not improve performance. In fact, peak performance is achieved when polysemic information is *ignored*, i.e., when all different annotated senses are collapsed to a single word, as they naturally appear in a text. These counter-intuitive results indicate that the word similarity tasks are not a suitable test to evaluate polysemic representations.

Although these negative results point to the inadequacy of the evaluation test, they also stress a necessary critical analysis of the sense-specific vectors. Specifically, they give rise to the following question: why might previously reported polysemic representations show superior performance in these inadequate evaluation tests?

One explanation for the reported effects may lie in an inherent property of sense-specific representations. The procedure of assigning a word occurrence to a particular sense amounts to a sampling procedure. This sampling procedure itself, regardless of its validity (whether its sense assignments are correct or not), may be the true source of the reported performance gains. To test this hypothesis, we created a control condition in which word occurrences are randomly assigned to different senses. Determining that an effect is attributable to *genuine polysemy* can only be established if a similar effect is lacking or significantly reduced in this control condition.

We demonstrate that performance gains are indeed obtained for a corpus with randomly assigned senses. In addition, we modify two models for sense-specific polysemy representation (Li and Jurafsky, 2015; Mancini et al., 2017) to randomly assign words to senses, and observe that the effect size remains unchanged between the original and random conditions.

In support of our empirical findings, we discuss the difficulty of obtaining an unbiased estimator for the cosine distance between two normalized random variable vectors. This distance is the basis of all word similarity tasks that serve to evaluate performance, and thus it may provide a partial explanation for the empirical findings, under the assumption that words are better represented as a population of vectors. Specifically, the true source of the reported performance gains may be an artifact of a purely statistical benefit that derives from the assignment of words to particular senses, or separate sub-samples, which subsequently reduces the bias of the similarity estimator.

We thus identify two independent pitfalls in NLP research on polysemy representation. First, the inadequacy of currently-used evaluation tests to properly assess polysemic representations. And second, the essence of polysemic representations, whose reported benefits might not come from polysemic information per se, but rather from other unrelated sources.

2 Background

Previous attempts to use polysemic information for enriching word representation used normal unannotated corpora, and therefore disambiguated the different senses of words before exploiting any sense-specific information. Prior art has either taken an automatic approach to detect word senses based on the statistics extracted from texts (Huang et al., 2012; Neelakantan et al., 2014; Li and Jurafsky, 2015), or relied on external lexical resources (e.g., WordNet or BabelNet) which guarantee that the detected senses are mapped to true sense distinctions in natural language (Chen et al., 2014; Iacobacci et al., 2015; Mancini et al., 2017). Importantly, both types of models report marked performance gains in evaluation tests.

This kind of approach produces (i) global vectors that represent a word's meaning as a single vector (with no subdivision into distinct senses), as well as (ii) sense-specific vectors representing individual senses of words, determined in the disambiguation step, as separate vectors. For example, such approaches would represent the meaning of *chair* as a single vector, as well as distinct vectors for each of its multiple senses, e.g., ”chair (person)” and ”chair (furniture)”.

In order to evaluate performance, the vectors created by the models are standardly evaluated using word similarity tasks (the most common are WordSim-353 (Finkelstein et al., 2001) and Stanford's Contextual Word Similarities (SCWS) (Huang et al., 2012)). These tasks comprise pairs of words and the similarity scores assigned to them by human annotators. For example, the similarity between *table* and *chair* might be rated as 0.8 (i.e., human annotators found these words to be very similar, but not perfectly so), while the similarity between *table* and *tree* might be rated as 0.3 (not very similar). The embedding models produce similarity scores for each word pair by computing the cosine-distance between the word vectors for each pair. The model's performance is then evaluated as the rank-order similarity in the order of pairs (Spearman correlation) between the human annotators' scores and the scores produced by the model. In line with the assumption discussed above, one would predict that the rank-order similarity produced by the sense-specific vectors should outperform the one produced by the global vectors. In particular, more accurate sense-specific vectors should produce better results in these tasks; conversely, better performance on these tasks is interpreted as indicating that polysemy has been captured more accurately. We directly test these two predictions in this paper.

Computing word similarity is straightforward when each word is represented as a single vector, but it is less so when the meaning of a polysemic word is represented by multiple sense-specific vectors. This problem of *matching* the senses relevant for a specific word pair, i.e., matching the "person" sense of *chair* with the correct sense of the word *meeting*, poses a major hurdle for meaningful comparison, and has been tackled in three different ways: (i) *average* over all similarity scores between all the different possible pairs; (ii) *weighted average* over these scores according to the probability of senses assigned by the disambiguation model; or (iii) *selection* of the most suitable sense according to the disambiguation model, and using only the corresponding similarity score.

Intuitively, the third approach should outperform the others, as it is based on the clearest distinction between the relevant and non-relevant senses. However, this naïve prediction is not supported by previous studies. Rather, the best results are usually obtained for *average* and *weighted average*, followed by *global* (ignoring polysemy), while *selection* falls far behind the others. This counter-intuitive finding suggests that the observed benefit may be less related to sense disambiguation than previously supposed[1].

3 Task validation

Generally, before any task can be used as an evaluation testbed for polysemy discovery algorithms or polysemous representations, we argue that the *task* itself should be validated as suitable (or not) for the intended purpose. We propose the following **task validation** methodology: (i) Start by identifying a corpus where polysemic information is known for a significant number of words. (ii) Compute two sets of word representations: $\mathcal{A}_1$ - which computes a single representation for all words in the corpus, and $\mathcal{A}_2$ - which computes multiple representations for each polysemic word in the corpus based on the different *known* senses of the word. (iii) Evaluate the task using the two representation sets $\mathcal{A}_1$ and $\mathcal{A}_2$. Only if significant performance gains can be shown when using $\mathcal{A}_2$ as compared to $\mathcal{A}_1$, the task can be used to evaluate polysemy representation.

3.1 Polysemy induction

A major drawback of the proposed methodology is that such annotated corpora are scarce, and the largest among them are still small (OntoNotes (Weischedel et al., 2013) comprises 1.5 million words, cf. unannotated corpora (e.g. Wikipedia) which are about 1000 times larger). We therefore articulate a methodology to generate a *task validation test* from **any corpus** and **evaluation task**, even without prior annotation of polysemy, that is based on the *pseudo-words* approach (Gale et al., 1992; Pilehvar and Navigli, 2014).

More specifically, we induce polysemy in a natural corpus by pairing words, and collapsing every pair of words into a single word-form while keeping their "original identity" as *polysemy annotation*. For example *ring* and *table* may be collapsed to a single word with two senses, *table*^{1} and *table*^{2} respectively. The new corpus is polysemic with respect to the collapsed words, while all other words keep a single sense. This corpus has most of the features of a natural corpus, but unlike most natural corpora (and all large corpora), it contains polysemy annotation. Subsequently, the relevant items in the *word similarity tasks* are collapsed in the same way, making the items polysemous, and thus the tasks suitable as *validation tests*.

With these polysemy-induced corpus and word similarity tasks, we follow the methodology for *task validation* described above. Only if a model that is based on multiple representations per polysemous word leads to a significant performance gain in the task as compared to a model with single representation for each word then the task un-

[1] Iacobacci et al. (2015) and Mancini et al. (2017) only reported results on *selection*, for which they found performance gains. For consistent comparison with other models we report *average* (but using their code), noting that it also provides performance gains over *global* vectors.

der examination should be considered adequate to evaluate the utility of polysemy representation of word senses.

3.2 Methods

Word embedding model *word2vec-SkipGram* model (Mikolov et al., 2013) is used to obtain vector representations for words and senses. The model is separately trained over the corpus: first producing sense-specific vectors according to the annotated senses, and next producing global vectors by ignoring the annotated senses and collapsing all their occurrences to a single word. Throughout the analyses we use an embedding size of 300d, a window size of 5 words from each side of the target word, negative-sampling of 5 words, and an initial learning rate of 0.025.

Sense-annotated corpus We use OntoNotes (Weischedel et al., 2013), the largest available corpus annotated for word senses. This allows us to circumvent the problem of first disambiguating the words' senses, and thus to directly test the utility of using polysemic information in word vector representations. The corpus contains 1.5 million English tokens, comprising about 50k English word types, of which 8675 word types are sense-annotated. Because the annotation is not uniform throughout the corpus (words are not annotated every time they appear), which can bias the analysis described below, we extract a subset of the corpus by removing sentences where polysemous words are not annotated, thus removing 40% of the corpus. Stopwords as well as words occurring less than 10 times are ignored by the word embedding models. All words are lowercased.

Sense-induced corpus Wikipedia dump (04/2017) is the original corpus from which a polysemic version is induced. We separately used a list of semantically-aware pseudowords (Pilehvar and Navigli, 2013), in addition to random pairs from the 6000 most frequent words to collapse them into a single word-form (see Section 3.1). Stopwords and infrequent words (<300 tokens) are ignored by the word embedding models.

Evaluation Polysemic word representations are evaluated on the two word similarity tasks described. Crucially, the problem of *matching* the relevant sense in these tests (described in Section 2) is tackled by taking the *average* of the sense-specific representations and comparing it to the *global* word representations[2].

3.3 Results

Results clearly demonstrate that global representations are significantly superior to sense-specific representations in both evaluation tests and across corpora, as shown in Table 1 and Figure 1.

	OPTIMAL	GLOBAL	AVERAGE
WS-353			
OntoNotes		44.7	41.3
Induced Rand.	70	68.7	66.4
Induced S.A.	70	70.2	67.1
SCWS			
OntoNotes		64.0	62.6
Induced Rand.	66	64.3	55.2
Induced S.A.	66	65.9	65.7

Table 1: Word similarity scores for OntoNotes and induced-polysemy Wikipedia. OPTIMAL: performance obtained for the original Wikipedia (apply only for the induced method), GLOBAL: sense information ignored, AVERAGE: sense-specific vectors averaged. S.A. & Rand.: pairing methods for polysemy induction, semantically-aware and random, respectively.

The critical comparison between the *global* and *average* conditions is expected to show better performance for the latter under the assumption that polysemic information improves performance. This difference (marked as *actual gain* in Figure 1), surprisingly shows an opposite outcome, which means that word similarity tasks fail to demonstrate the value of polysemy representation in improving performance.

Complementarily, we report performance for the *optimal* condition. In this condition, word similarity scores are obtained for a model that was trained on the the original Wikipedia corpus (before polysemy induction), and thus represents an upper bound for the performance of any polysemy model that would be trained on an induced polysemy version of the same corpus (i.e., the maximum *ideal gain* a model can achieve over its *global* vectors). The fact that performance is highest for that condition is thus expected, and reassures us that the induced polysemy procedure has worked as planned for both kinds of induction.

Overall, our results points to a failure of polysemy models to improve performance over *global* vectors by *averaging* sense-specific vectors. This worsening in performance of the sense-specific

[2] Recall that *average* was reported to be one of the best performing matching methods is previous work.

vectors stands in marked contrast to previous studies which reported performance gains (see Table 2 and *Reported gains* in Figure 1).

	GLOBAL	AVERAGE
WS-353		
Huang (2012)	22.8	71.3
Neelakantan (2014)	69.2	70.9
Li (2015)*	61.0	67.8
Mancini (2017)*	49.1	55.6
SCWS		
Huang (2012)	58.6	62.8
Neelakantan (2014)	65.5	67.2
Chen (2014)	64.2	66.2
Li (2015)	64.6	66.4
Mancini (2017) *	57.2	62.1

Table 2: Previously reported results in word similarity tasks. *Reproduced (see section 4.1) where consistent comparison on the same tasks is not originally reported.

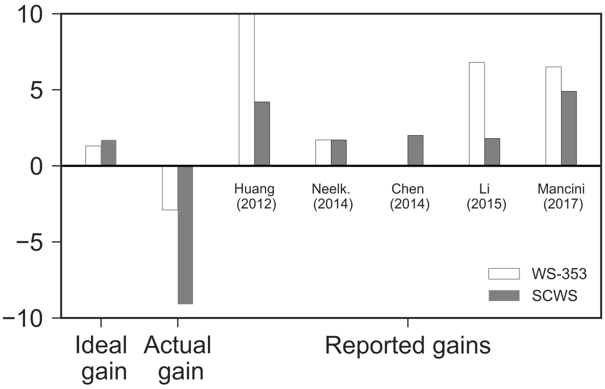

Figure 1: Summary of results in Tables 1, 2, showing *performance gains* across conditions. *Ideal gain*: max. performance gains using fully annotated corpus with perfect sense matching, *Actual gain*: de-facto gain (global - average), and *Reported gains*: previous reported results. Color (dark and white) marks the tasks.

3.4 Discussion

The main result described above is negative, demonstrating that *word similarity tasks* are not suitable to evaluate polysemy representation. However, the methodology we proposed for *polysemy induction* constitutes a positive contribution, as it can be used to effectively test any task for its utility in the evaluation of polysemy representation while using any corpora. This may aid in finding tasks which are more suitable to serve as gold standard evaluation tests for polysemy. Moreover, the use of polysemy induction for these purposes adds yet another type of control to the NLP toolbox; such controls are scarcely implemented in NLP studies (but see Dubossarsky et al. (2017)).

Importantly, the inability of the induced polysemy models to produce positive performance gains as reported in prior art may indicate that these reported gains do not reflect benefits from true polysemy, but rather from an unknown factor that boosts performance.

4 The statistical signature of polysemy

In order to further investigate the reason for the lack of performance gains, we analyzed the properties of the sense-specific vectors in the induced polysemy conditions, and compared them to those obtained by previous studies.

We looked at the pairwise similarity between the different sense-specific representations of the same word. We start from the observation that polysemy is inherently defined by word senses that are distinguishable from each other. Importantly, we compared the models that did not produce performance gains in word similarity tasks to those which did report such gains.

4.1 Analysis and results

We tested four sets of sense-specific representations: two from our polysemy induction models (see Section 3.2) and two from Li and Jurafsky (2015)[3] and Mancini et al. (2017)[4] which reported performance gains. For each word in each set, the average cosine-distance between its different sense-specific vectors is computed. The distribution of these average distances within a specific set is defined as its ”polysemic signature”, which is then compared across sets.

The results shown in Figure 2 reveal marked differences in the polysemic signature between the four sets. A high polysemic signature is seen for the two induced polysemy sets, which are the only sets with guaranteed semantically different senses. In these sets, the polysemy model that is based on the random pairing of words has a higher polysemic signature than the one based on semantically-aware pairing. This is well expected, as semantically-aware pseudowords are designed to simulate ”true polysemy” in which the different senses of a word are still semantically related, unlike random pairing which produces a ”coarse” distinction more similar to homonymy (Pilehvar and Navigli, 2014).

The critical comparison to the two sets of polysemy models that did find performance gains

[3] https://github.com/jiweil/mutli-sense-embedding

[4] http://lcl.uniroma1.it/sw2v/

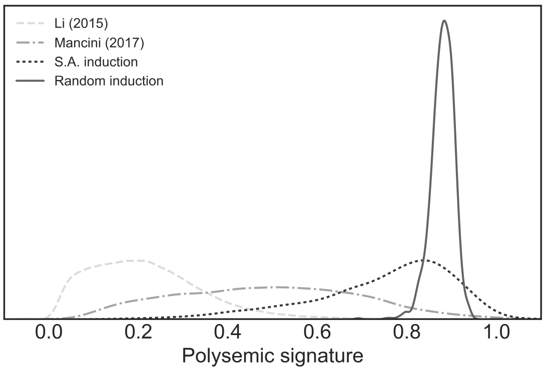

Figure 2: Density distribution of polysemic signatures for the four sets, see text for details.

shows that these models exhibit a smaller polysemic signature, as the different senses seem to be more similar to one another. Crucially, even the polysemic signature of Mancini et al. (2017), which exhibit an intermediate polysemic signature, shows greater similarity in its senses as compared to the sense-aware polysemy induction model, which presumably represents a more subtle (and ecological) model for polysemy.

4.2 Discussion

The broader implications of these results on our research hypothesis can be understood in the context of the findings reported in Section 3.3. The results described in Figure 1 demonstrate a marked difference between previously *reported gains* and the *actual gain* condition which shows a **worsening** of performance in the same task when polysemic information is included. The results demonstrated in Figure 2 can be described as a negative image of those presented in Figure 1. Specifically, the *actual gain* condition of the induced-polysemy has the largest polysemic signature as compared to the other conditions.

Together, these results indicate that the condition that demonstrates polysemy most clearly shows the poorest performance in the evaluation tests. The converse is also true: the conditions that diverge from the clearest polysemic representation show heightened performance in the evaluation tests. A gold standard for polysemy representation should entail that given optimal vector representations, performance on the evaluation tests would be optimal, and vice versa. Since our results demonstrate that the directions of optimal vector representation and optimal test performance are opposite, we are led to the following conclusions: First, polysemy models which provide improvement in word similarity tasks may not necessarily be suitable for polysemous vector representations. Second, word similarity tasks are not suitable evaluation tests for polysemy representation.

5 Theoretical discussion

In this section we recall and analyze some properties of the cosine distance, and describe how they may partially account for the empirical observations discussed in this article. The crucial point is to model the contextual representation of words as a distribution over some vector space.

Let X_i denote the random variable which captures the contextual representation of word i. Let $\{X_i^l, X_j^l \in \mathbb{R}^d\}_{l=1}^L$ denote a sample of such representations for words i, j respectively, where L denotes the sample size. d corresponds to the dimension of the vector space when using word2vec representations, or the number of words in the dictionary when using explicit representations (e.g., PPMI) (see Section 3.2). To simplify the analysis, we further assume that $\|X^l\| = 1 \; \forall l$.

The similarity between two words i, j can be plausibly measured (as customarily done) by the cosine distance between their contextual representations, namely, $\mathbb{E}[X_i X_j]$:

$$\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\,\mathbb{E}[X_j] + \operatorname{cov}(X_i, X_j) \quad (1)$$

Thus the average distance is not equivalent to the distance between the average representations alone, but has to include an additional bias term - $\operatorname{cov}(X_i, X_j)$ - which reflects the statistical dependence between the two vector representations X_i, X_j. This term is significant, because the contextual representation of two words is likely to exhibit strong dependence, especially when the words are more similar.

This is where the problem lies. In the process of generating words' representations, we start from a sample of sentences and generate a single representation. This representation is essentially our estimate of $\mathbb{E}[X_i]$ for word i. When multiplying two such representations in order to compute the cosine distance between them, we obtain an estimate for $\mathbb{E}[X_i]\,\mathbb{E}[X_j]$, which is *not* a good estimate for $\mathbb{E}[X_i X_j]$ because of the bias term in (1).

Ideally, in order to provide an unbiased estimate of $\mathbb{E}[X_i X_j]$, we should divide the sample of sentences into mini-batches, compute the appropriate contextual representation for both words

i, j from each mini-batch, and then directly estimate $\mathbb{E}\left[X_i X_j\right]$ by taking the average multiplication of the corresponding representations in each mini-batch. Interestingly, in the process of generating polysemous representations, whether relying on true polysemy or arbitrary polysemy, we essentially accomplish the same goal: for each word, a mini-batch is replaced by the subset of sentences in which only one of the word's meanings is present[5].

If sense matching (see Section 2) is achieved by way of *average* or *weighted average*, it implies that our estimate of word similarity should improve with the number of senses used in the analysis, especially when the assignment is arbitrary. Of course, any improvement is hampered by the deterioration in the quality of the contextual representation computed from the smaller mini-batch sample, and therefore improvement is only expected for a small number of real or artificial ”senses”.

6 Performance gains revisited

The empirical findings presented so far converge on the conclusion that the performance gains reported in prior art may not stem from the utility of polysemic information, as previously claimed, but are the result of an alternative source. In the theoretical discussion we argue that random sense annotation is equivalent to sub-sampling and multiple estimation of contextual vector representations, and that this alone may be beneficial for the task performance of word similarity. It is reasonable to conclude that these repeated sampling procedures may have produced the reported performance gains. Here we test this hypothesis directly.

We propose a simple control condition, in which senses are randomly assigned to words in a corpus, and sense-specific vectors are produced in the same way as before. Determining that an effect is reliably attributed to *genuine polysemy* can only be established if a similar effect is lacking or significantly reduced in this control condition.

6.1 Random sense assignment

We achieve random sense assignments in 2 ways:

Sampling from a known distribution. For the entire corpus and vocabulary (100k words, and note that Neelakantan et al. (2014) and Li and Jurafsky (2015) also took this entire vocabulary approach), we assigned senses at random from a categorical distribution under two conditions. In the first we used a uniform prior which produced equal sense probabilities for each word, and in the second condition we used a biased prior in which one sense predominates. We also experimented with different number of senses per word. We found that the results differed only slightly between the conditions and across the different number of senses. The same regular embedding model was trained as before.

Sampling from an unknown distribution. To test the hypothesis more directly against the sense distributions used in prior work, we reproduced sense-specific vectors of two models using their code: (1) for Li and Jurafsky (2015) we kept their Chinese-Restaurant-Process probabilistic mechanism, where senses are assigned to words based on the similarity of their contexts. We only shuffled the elements of the final vector of sense assignments produced by the model, and (2) for Mancini et al. (2017) we shuffled between the sense tags of each word in their original sense-annotated Wikipedia. For further comparisons we used the original code unchanged to reproduce another set of global and sense-specific vectors for each model separately to serve as a baseline.

6.2 Performance boost due to word sampling

Table 3 shows the results of these simulations. Regular embedding with random sense assignments shows marked performance gains of the sense-specific vectors over the global vectors. In fact, the effect reported in Neelakantan et al. (2014) (which report the highest score in the SCWS task, see Table 2) is replicated almost exactly, perhaps due to the fact that they also used a fixed number of senses for each word as we did in this simulation. Furthermore, the reanalysis of previous models lead to almost identical results in the original and random conditions, which means that randomly assigning words to senses does not weaken the effect.

Together, these three control simulations clearly show that an effect of the same magnitude as previously reported in several studies emerges under random sense assignment. Therefore, our findings strongly undermine the assumption that the reported effects are in any way related to actual polysemy, and strongly suggests that it is repeated sub-sampling that boosts performance.

[5] For the purpose of this discussion we ignore sentences in which a word appears more than once.

	ORIGINAL		RANDOM	
	GLOB.	AVE.	GLOB.	AVE.
WS-353				
Regular embd			70	69.8
Li (2015)	61.0	67.8	60.5	68.2
Mancini (2017)	49.1	55.6	58.3	64.5
SCWS				
Regular embd			65.7	67.3
Li (2015)	58.9	66.2	57.4	66.2
Mancini (2017)	57.2	62.1	62.4	65.2

Table 3: Word similarity scores for random sense assignments, compared to the original (where available).

7 Summary and Discussion

Here we investigate the common-held view that polysemy representation improves performance. Specifically, we question the validity of word similarity tasks as a suitable evaluation test that would allow drawing such conclusions. To test the claim that resolving the polysemy of words improves performance in these tasks, as was repeatedly reported in prior art, we used real-world polysemy in two independent conditions: (i) a human sense-annotated corpus and (ii) a corpus in which polysemy was induced in a controlled artificial fashion. In both conditions, the performance in word similarity tasks *deteriorated.* We claim that this negative finding alone may suffice to determine that word similarity tasks are not suitable tests for evaluating polysemy representation.

However, if the word similarity task is inadequate to evaluate polysemy, why would it show high gains for polysemic representations as reported in many previous studies? To investigate this question we first ask whether polysemic information per se is required to drive such effects, or whether these effects are artificially caused by the procedures by which polysemous representations are created. To test this, we set out to demonstrate that even representations that bear no genuine polysemic information could nonetheless yield improved performance due to a methodological artifact. We thus created control conditions, in which we randomly assign word occurrences to senses, and found that randomly-produced sense-specific vectors indeed show marked improvements in performance. Since this effect cannot stem from polysemy (which is lacking in these conditions), it may only be the result of a methodological artifact - the sampling procedure entailed by the assignment of words to senses.

The existence of a sampling artifact is supported by our theoretical discussion, showing that multiple vector sampling can lower the bias of the estimator of the cosine distance between two vectorial random variables. The underlying assumption is that a better model for contextual word representation should employ a population of vectors. Interestingly, the conclusion that word vectors are better if constructed from multiple representations might apply to word vectors in general, and not just to sense-specific vectors. However, such a claim is outside the scope of this study and remains for future research.

This account is further supported by our *polysemic signature* analysis, which shows that the similarity between the senses in the polysemic models that produced performance gains is greater than in the models that did not produce this effect (the polysemy-induced models). This finding is in line with the sampling artifact account, as sense-specific vectors which are based on random sense assignments are expected to be more similar compared to sense-specific vectors which are based on true polysemic distinctions.

We stress that our analysis does not argue for or against the accuracy of sense-specific vectors in capturing true polysemy (other tests exist for that purpose). Instead, it focuses only on the true source of previously reported performance gains of this kind of representation, and on the validity of word similarity tasks for their evaluation.

All in all, we provide converging evidence, both experimental and theoretical, that word similarity tasks do not provide a marker for the utility of polysemic information. Rather, performance gains in word similarity tasks are an artifact of the procedure by which polysemic representations are created. These conclusions join a general skepticism expressed in the literature about the use of word similarity tasks for the evaluation of word vectors (Hill et al., 2015; Avraham and Goldberg, 2016; Batchkarov et al., 2016; Faruqui et al., 2016).

Essentially, our findings mean that there is no solid empirical foundation to the claim that polysemic information improves performance in evaluation tests. In fact, they corroborate the general impression that polysemic representation does not improve performance on most downstream tasks (Li and Jurafsky, 2015). It may be the case that sense-specific vectors can or will show heightened performance on evaluation tests, or improve downstream tasks. This, however, will have to be demonstrated on the basis of tasks whose suitability has been properly validated. Moreover, any effects reported will have to be supported by demonstrating that these effects are absent or strongly reduced in a properly articulated condition that should control for the sampling artifact.

Critically, until a reliable evaluation test exists, research on polysemic word representation is seriously hampered. In fact, a re-evaluation of past models would be in place, as both ”positive” and ”negative” results that were previously reported are in fact invalid.

Acknowledgments

We thank Matan Ben-Yosef for his invaluable contribution in executing and manipulating the code of Li and Jurafsky (2015).

References

Oded Avraham and Yoav Goldberg. 2016. Improving reliability of word similarity evaluation by redesigning annotation task and performance measure. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 106–110.

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

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of EMNLP*, pages 1025–1035.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of EMNLP*, pages 1136–1145.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. Work on statistical methods for word sense disambiguation. In *AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, volume 54, page 60.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y 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: Long Papers-Volume 1*, pages 873–882. 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)*, volume 1, pages 95–105.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *Proceedings EMNLP*, pages 1722–1732.

Massimiliano Mancini, Jose Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. 2017. Embedding words and senses together via joint knowledge-enhanced training. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 100–111.

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

Arvind Neelakantan, Jeevan Shankar, Re Passos, and Andrew Mccallum. 2014. Efficient nonparametric estimation of multiple embeddings per word in vector space. In *Proceedings of EMNLP*.

Mohammad Taher Pilehvar and Roberto Navigli. 2013. Paving the way to a large-scale pseudosense-annotated dataset. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1100–1109.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation. *Computational Linguistics*, 40(4):837–881.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*.