When the Whole is Less than the Sum of its Parts: How Composition Affects PMI Values in Distributional Semantic Vectors

Denis Paperno and Marco Baroni
University of Trento

Distributional semantic models, deriving vector-based word representations from patterns of word usage in corpora, have many useful applications (Turney and Pantel 2010). Recently, there has been interest in compositional *distributional models, which derive vectors for phrases from representations of their constituent words (Mitchell and Lapata 2010). Often, the values of distributional vectors are Pointwise Mutual Information (PMI) scores obtained from raw co-occurrence counts. In this paper we study the relation between the PMI dimensions of a phrase vector and its components in order to gain insights into which operations an adequate composition model should perform. We show mathematically that the difference between the PMI dimension of a phrase vector and the sum of PMIs in the corresponding dimensions of the phrases' parts is an independently interpretable value, namely a quantification of the impact of the context associated to the relevant dimension on the phrase's internal cohesion, as also measured by PMI. We then explore this quantity empirically, through an analysis of adjective-noun composition.*

1 Introduction

Pointwise Mutual Information Dimensions of a word vector in distributional semantic models contain a function of the co-occurrence counts of the word with contexts of interest. A popular and effective option (Bullinaria and Levy 2012) is to transform counts into PMI scores, given, for any word a and context c, by $PMI(a, c) = \log(\frac{P(a|c)}{P(a)})$.[1]
Composition methods There are various proposals on deriving phrase representations by composing word vectors, ranging from simple, parameter-free vector addition to fully supervised deep-neural-network-based systems. We focus here on the models illustrated in Table 1; see Dinu, Pham, and Baroni (2013) for the original model references. As an empirical test case, we consider adjective-noun composition.

1 PMI is also used to measure the tendency of two phrase constituents to be combined in a particular syntactic configuration (e.g., to assess the degree of lexicalization of the phrase). We use $PMI(ab)$ to refer to this "phrase cohesion" PMI. $PMI(ab) = \log(\frac{P(ab)}{P(a)\cdot P(b)}) = \log(\frac{P(ab|a)}{P(b)})$ depends on how often words a and b form a phrase of the relevant kind, e.g., how often adjective a modifies noun b, while $PMI(a, b)$ is based on the probability of two words to occur in the relevant co-occurrence context (e.g., within an n-word window). Quantifying phrase cohesion with $PMI(ab)$ has been the most common usage of PMI in computational linguistics outside distributional semantics ever since the seminal work on collocations of Church and Hanks (1990).

model	*phrase vector*	*model*	*phrase vector*
additive	$\vec{a} + \vec{b}$	full additive	$\mathbf{X}\vec{a} + \mathbf{Y}\vec{b}$
multiplicative	$\vec{a} \odot \vec{b}$	lexical function	$\mathbf{A}\vec{b}$
weighted additive	$\alpha\vec{a} + \beta\vec{b}$	shifted additive	$\vec{a} + \vec{b} + \vec{c}$

Table 1
Composition methods. $\odot$ is pointwise multiplication. α and β are scalar parameters, matrices $\mathbf{X}$ and $\mathbf{Y}$ represent syntactic relation slots (e.g., Adjective and Noun), matrix $\mathbf{A}$ represent a functional word (e.g., the adjective in an adjective-noun construction) and $\vec{c}$ is a constant vector.

2 A general result on the PMI dimensions of phrases

An ideal composition model should be able to reconstruct, at least for sufficiently frequent phrases, the corpus-extracted vector of the phrase ab from vectors of its parts a, b. When vector dimensions encode PMI values, for each context c, the composition model has to predict PMI(ab, c) between phrase ab and context c. Eq. (1) shows that there is a mathematical relation between PMI(ab, c) and the PMI values of the phrase components PMI(a, c), PMI(b, c):

$$\begin{aligned} PMI(ab,c) &= \log(\frac{P(ab \mid c)}{P(ab)}) = \log(\frac{P(a \mid c) \cdot P(ab \mid a \wedge c)}{P(a) \cdot P(ab \mid a)}) = \\ &= \log(\frac{P(a \mid c) \cdot P(ab \mid a \wedge c)}{P(a) \cdot P(ab \mid a)} \cdot \frac{P(b \mid c) \cdot P(b)}{P(b \mid c) \cdot P(b)}) = \\ &= \log(\frac{P(a \mid c)}{P(a)} \cdot \frac{P(b \mid c)}{P(b)} \cdot \frac{P(ab \mid a \wedge c)}{P(b \mid c)} \cdot \frac{P(b)}{P(ab \mid a)}) = \\ = \log(\frac{P(a \mid c)}{P(a)}) + \log(\frac{P(b \mid c)}{P(b)}) + \log(\frac{P(ab \mid a \wedge c)}{P(b \mid c)}) - \log(\frac{P(ab \mid a)}{P(b)}) &= \\ = PMI(a,c) + PMI(b,c) + PMI(ab|c) - PMI(ab) & \end{aligned} \tag{1}$$

To make sense of this derivation, observe that $P(ab)$ and $P(ab \mid c)$ pertain to a phrase ab where a and b are linked by a specific syntactic relation. Now, whenever the phrase ab occurs, a must also occur, and thus $P(ab) = P(ab \wedge a)$, and similarly $P(ab \mid c) = P(ab \wedge a \mid c)$. This connects the PMI of a phrase (based on counts of ab linked by a syntactic relation) to the PMI of the constituents (based on counts of the constituents in all contexts). Consequently, we can meaningfully relate $PMI(ab, c)$ (as computed to calculate phrase vector dimensions) to $PMI(a, c)$ and $PMI(b, c)$ (as computed to calculate single word dimensions).

Eq. (1) unveils a systematic relation between the PMI value in a phrase vector dimension and the value predicted by the *additive* approach to composition. Indeed, $PMI(ab, c)$ equals $PMI(a, c) + PMI(b, c)$, shifted by some correction $\Delta_c PMI(ab) = PMI(ab|c) - PMI(ab)$, measuring how the context changes the tendency of two words a, b to form a phrase. Δ_c includes any non-trivial effects of composition arising from the interaction between the occurrence of words a, b, c. Absence of non-trivial interaction of this kind is a reasonable null hypothesis, under which the association of phrase components with each other is not affected by context at all: $PMI(ab|c) = PMI(ab)$. Under this null hypothesis, addition should accurately predict PMI values for phrases.

3 Empirical observations

We have shown that vector addition should perfectly predict phrase vectors under the idealized assumption that the context's effect on the association between words in the phrase, $\Delta_c PMI(ab) = PMI(ab|c) - PMI(ab)$, is negligible. $\Delta_c PMI(ab)$ equals the deviation of the actual $PMI(ab, c)$ from the additive ideal, which any vector composition model is essentially trying to estimate. Let us now investigate how well actual vectors of English phrases fit the additive ideal, and, if they do not fit, how good the existing composition methods are at predicting deviations from the ideal.

Experimental setup We focus on adjective-noun (AN) phrases as a representative case. We used 2.8 billion tokens comprised of ukWaC, Wackypedia and British National Corpus,[2] extracting the 12.6K ANs that occurred at least 1K times. We collected sentence-internal co-occurrence counts with the 872 nouns[3] occurring at least 150K times in the corpus used as contexts. PMI values were computed by standard maximum-likelihood estimation.

We separated a random subset of 6K ANs to train composition models. We consider two versions of the corresponding constituent vectors as input to composition: plain PMI vectors (with zero co-occurrence rates conventionally converted to 0 instead of $-\infty$) and PPMI vectors (all non-positive PMI values converted to 0). The latter transformation is common in the literature. Model parameters were estimated using DISSECT (Dinu, Pham, and Baroni 2013), whose training objective is to approximate corpus-extracted phrase vectors, a criterion especially appropriate for our purposes.

We report results based on the 1.8 mln positive PMI dimensions of the 4.7K phrase vectors that were not used for training.[4] On average a phrase had non-zero co-occurrence with 84.8% of context nouns, over half of which gave positive PMI values. We focus on positive dimensions because negative association values are harder to interpret; furthermore, $-\infty$ cases must be set to some arbitrary value, and most practical applications set all negative values to 0 anyway (PPMI). We also repeated the experiments including negative observed values, with a similar pattern of results.

Divergence from additive We first verify how the observed PMI values of phrases depart from those predicted by addition: in other words, how much $\Delta_c PMI(ab) \neq 0$ in practice. We observe that $PMI(ab, c)$ has a strong tendency to be *lower* than the sum of PMI of the phrase's parts wrt the same context. In our sample, average $PMI(AN, c)$ was .80, while average PMI(A, c) and PMI(N, c) were .55 and .63, respectively.[5] Over 70% of positive PMI values in our sample are lower than additive ($PMI(AN, c) < PMI(A, c) + PMI(N, c)$); a vast majority of phrases (over 92%) have on average a negative divergence from the additive prediction, $\frac{\sum_{c \in C} PMI(AN,c) - (PMI(A,c) + PMI(N,c))}{|C|} < 0$. The tendency for phrases to have lower PMI than predicted by the additive idealization is quite robust. It holds whether or not we restrict the data to items with positive PMI of constituent words ($PMI(A, c) > 0, PMI(N, c) > 0$), if we convert all negative PMI

2 `http://wacky.sslmit.unibo.it/`, `http://www.natcorp.ox.ac.uk/`

3 Only nouns were used to avoid adding the context word's part of speech as a parameter of the analysis. The number of contexts used was restricted by the consideration that training the lexical function model for larger dimensionalities is problematic.

4 About 1.9K ANs containing adjectives occurring with less than 5 context nouns were removed from the test set at this point, because we would not have had enough data to train the corresponding lexical function model for those adjectives.

5 We cannot claim this divergence on unattested phrase-context co-occurrences because those should give rise to very small, probably negative, PMI values.

values of constituents to 0, and also if we extend the test set to include negative PMI values of phrases ($PMI(AN, c) < 0$).

A possible reason for the mostly negative deviation from addition comes from the information-theoretic nature of PMI. Recall that $PMI(ab)$ measures how informative phrase components a, b are about each other. The negative deviation from addition $\Delta_c PMI(ab)$ means that context is diminishing the mutual information of a and b. And indeed it is only natural that the context itself is usually informative. Concretely, it can be informative in multiple ways. In one typical scenario, the two words being composed (and the phrase) share the context topic (e.g., *logical* and *operator* in the context of *calculus*, connected by the topic of mathematical logic). In this case there is little additional PMI gained by composing such words since they share a large amount of co-occurring contexts. Take the idealized case when the *shared underlying topic* increases the probability of A, N, and AN by some constant k, so $PMI(A, c) = PMI(N, c) = PMI(AN, c) = \log k$. Then association (PMI) of AN decreases by $\log k$ in the presence of topic-related words, $\Delta_c PMI(AN) = PMI(AN, c) - (PMI(A, c) + PMI(N, c)) = -\log k$. The opposite case of negative association between context and A,N is not symmetric to the positive association just discussed (if it were, it would have produced a positive deviation from the additive model). Negative association is in general less pronounced than positive association: in our sample, positive PMI values cover over half the co-occurrence table; furthermore positive PMIs are on average greater in absolute value than negative ones. Importantly, two words in a phrase will often disambiguate each other, making the phrase less probable in a given context than expected from the probabilities of its parts: *logical operator* is very unlikely in the context of *automobile* even though *operator* in the sense of a person operating a machine and *logical* in the non-technical sense are perfectly plausible in the same context. Such disambiguation cases, we believe, largely account for negative deviation from additive in the case of negative components.

One can think of minimal adjustments to the additive model correcting for systematic PMI overestimation. Here, we experiment with a *shifted additive* model obtained by subtracting a constant vector from the summed PMI vector. Specifically, we obtained shifted vectors by computing, for each dimension, the average deviation from the additive model in the training data.

Approximation to empirical phrase PMI by composition models We have seen that addition would be a reasonable approximation to PMI vector composition if the influence of context on the association between parts of the phrase turned out to be negligible. Empirically, phrase-context PMI is systematically negatively deviating from word-context PMI addition. Crucially, an adequate vector composition method should capture this deviation from the additive ideal. The next step is to test existing vector composition models on how well they achieve this goal.

To assess approximation quality, we compare the $PMI(AN, c)$ values predicted by each composition model to the ones directly derived from the corpus, using mean squared error (MSE) as figure of merit. Besides the full test set (*all* in Table 2), we consider some informative subsets. The *pos* subset includes the 40K *AN,c* pairs with largest positive error wrt the additive prediction (above 1.264). The *neg* subset includes the 40K dimensions with the largest negative error wrt additive (under -1.987). Finally, the *near-0* subset includes the 20K items with the smallest positive errors and the 20K items with the smallest negative errors wrt additive (between -0.026 and 0.023). Each of the three subsets constitutes about 2% of the *all* dataset.

By looking at Table 2, we observe first of all that addition's tendency to overestimate phrase PMI values puts it behind other models in the *all* and *neg* test sets, even behind the multiplicative method, which unlike others has no theoretical motivation.

	PMI				PPMI			
model	*all*	*pos*	*neg*	*near-0*	*all*	*pos*	*neg*	*near-0*
additive	0.75	3.11	5.86	(≈0.00)	0.71	1.96	5.88	(0.02)
multiplicative	0.61	2.50	3.38	0.55	0.59	5.92	3.40	0.62
weighted additive	0.39	2.52	**0.41**	0.35	**0.32**	3.01	0.62	0.27
full additive	0.56	3.02	0.68	0.54	0.34	**1.93**	0.63	0.29
lexical function	0.73	2.92	0.74	0.68	0.45	2.01	0.68	0.37
shifted additive	0.66	4.66	3.01	0.39	0.48	2.16	3.52	**0.18**

Table 2
MSE of different models' predictions, trained on PMI (left) vs. PPMI vectors (right).

The relatively good result of the multiplicative model can be explained through the patterns observed earlier: PMI(ab,c) is typically just above PMI(a,c) and PMI(b,c) for each of the phrase components (median values .66, .5, and .56, respectively). Adding PMI(a,c) and PMI(b,c) makes the prediction further above the observed PMI(ab,c) than their product is below it (when applied to median values, we get deviations of $|.66 - (.5 * .56)| = .38$ for multiplication and $|.66 - (.5 + .56)| = .4$ for addition). As one could expect, shifted addition is on average closer to actual PMI values than plain addition. However, weighted addition provides better approximations to the observed values. Shifted addition behaves too conservatively wrt addition, providing a good fit when observed PMI is close to additive (near-0 subset), but only bringing about a small improvement in the all-important negative subset. Weighted addition, on the other hand, brings about large improvements in approximating precisely the negative subset. Weighted addition is the best model overall, outperforming the parameter-rich full additive and lexical function models (the former only by a small margin). Confirming the effectiveness of the non-negative transform, PPMI-trained models are more accurate than PMI-trained ones, although the latter provide the best fit for the extreme negative subset, where component negative values are common.

As discussed above, the observed deviation from additive PMI is mostly negative, due partly to the shared underlying topic effect and partly to the disambiguation effect. In both cases, whenever the PMI of the constituents ($PMI(a, c)$ and/or $PMI(b, c)$) is larger, the deviation from additive ($PMI(ab, c) - (PMI(a, c) + PMI(b, c))$) is likely to become smaller. Weighted addition captures this, setting the negative correction of the additive model to be a linear function of the PMI values of the phrase components. The full additive model, which also showed competitive results overall, might perform better with more training data or with lower vector dimensionality (in the current setup, there were just about 3 training examples for each parameter to set).

4 Conclusion

We showed, based on the mathematical definition of PMI, that addition is a systematic component of PMI vector composition. The remaining component is also an interpretable value, measuring the impact of context on the phrase's internal PMI. In practice, this component is typically negative. Empirical observations about adjective-noun phrases show that systematic deviations from addition are largely accounted for by a negative shift $\Delta_c PMI(ab)$, which might be proportional to the composed vectors' dimensions (as partially captured by the weighted additive method). Further studies

should consider other constructions and types of context, to confirm the generality of our results.

Acknowledgments

We would like to thank the Computational Linguistics editor and reviewers, Yoav Goldberg, Omer Levy, Germán Kruszewski, Nghia Pham and the other members of the Composes team for useful feedback. Our work is funded by ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

References

Bullinaria, John and Joseph Levy. 2012. Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, 44:890–907.

Church, Kenneth and Peter Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Dinu, Georgiana, Nghia The Pham, and Marco Baroni. 2013. General estimation and evaluation of compositional distributional semantic models. In *Proceedings of ACL Workshop on Continuous Vector Space Models and their Compositionality*, pages 50–58, Sofia, Bulgaria.

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

Turney, Peter and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.