Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4379–4387
Marseille, 11–16 May 2020

Variants of Vector Space Reductions for Predicting the Compositionality of English Noun Compounds

Pegah Alipoor[1], **Sabine Schulte im Walde**[2]
[1]Sharif University of Technology, Islamic Republic of Iran
[2]Institute for Natural Language Processing, University of Stuttgart, Germany
[1]palipoor@ce.sharif.edu
[2]schulte@ims.uni-stuttgart.de

Abstract

Predicting the degree of compositionality of noun compounds such as *snowball* and *butterfly* is a crucial ingredient for lexicography and Natural Language Processing applications, to know whether the compound should be treated as a whole, or through its constituents, and what it means. Computational approaches for an automatic prediction typically represent and compare compounds and their constituents within a vector space and use distributional similarity as a proxy to predict the semantic relatedness between the compounds and their constituents as the compound's degree of compositionality. This paper provides a systematic evaluation of vector-space reduction variants across kinds, exploring reductions based on part-of-speech next to and also in combination with Principal Components Analysis using Singular Value Decomposition, and word2vec embeddings. We show that word2vec and nouns-only dimensionality reductions are the most successful and stable vector space reduction variants for our task.

Keywords: Noun Compounds, Compositionality, Vector Spaces, Dimensionality Reduction

1. Introduction

Predicting the degree of compositionality of noun compounds (and multi-word expressions in more general) is a crucial ingredient for lexicography and Natural Language Processing (NLP) applications, to know whether the expression should be treated as a whole, or through its constituents, and what the expression means. Compare, for example, the English noun compounds *snowball* –a ball consisting of snow, where clearly both constituents *snow* and *ball* contribute to the meaning of the compound– and *butterfly* –where the semantic contribution of the modifier noun *butter* is not obvious without knowing about the etymology of the compound. Studies such as Cholakov and Kordoni (2014), Weller et al. (2014), Cap et al. (2015), and Salehi et al. (2015b) are examples of NLP applications that have integrated the prediction of multi-word compositionality into statistical machine translation.

Accordingly, the field has witnessed a rich amount of computational approaches to automatically predict the degree of compositionality of noun compounds. These approaches typically represent compounds and their constituents within a vector space, and then compare the compound vectors with the constituent vectors as a proxy to the compounds' degree of compositionality (Reddy et al., 2011b; Reddy et al., 2011a; Salehi and Cook, 2013; Schulte im Walde et al., 2013; Salehi et al., 2014; Schulte im Walde et al., 2016; Cordeiro et al., 2019). Most of the approaches focus on English and German; most recently, Cordeiro et al. (2019) applied their framework to also French and Portuguese.

All of the above-mentioned approaches explored variants of vector space models in some way, regarding the composite functions to combine the constituent vectors (Reddy et al., 2011b); or regarding the translations of compounds and constituents into multiple languages (Salehi et al., 2014); or regarding the contributions of modifiers and heads (Schulte im Walde et al., 2016); etc. What is still lacking, however, is a systematic assessment of the effect of vector-space reductions on the quality of predicting compositionality: Bullinaria and Levy (2012) explored the effect of Singular Value Decomposition (SVD) on semantics in vector spaces in general; and from Baroni et al. (2014b) and Levy et al. (2015) –among many others– we know that word embeddings provide a useful low-dimensional representation for vector spaces. But as to our knowledge, up to date only Salehi et al. (2015a) and Cordeiro et al. (2019) integrated vector-space reductions (in the form of word embeddings) into their computational prediction of noun compound compositionality, and Schulte im Walde et al. (2013) explored part-of-speech-based reductions in combination with frequency effects.

Our contribution in this paper is to provide a systematic evaluation of vector-space reductions across kinds, i.e., exploring part-of-speech-based reduction, Principal Components Analysis using Singular Value Decomposition, and word2vec embeddings. Relying on the English noun compound dataset by Reddy et al. (2011b) as our gold standard, we show that word2vec and nouns-only dimensionality reductions are the most successful and stable vector space variants for our task.

2. Related Work

Most closely related studies includes distributional approaches that predict the degree of compositionality of a compound regarding a specific constituent (by comparing the compound vector to the respective constituent vector), or a functional combination of several constituents' vectors. Most importantly, Reddy et al. (2011b) used a standard distributional model to predict the compositionality of compound-constituent pairs for 90 English compounds. They extended their predictions by applying composite functions (see above). In a similar vein, Schulte im Walde et al. (2013) predicted the compositionality for 244

German compounds, and Schulte im Walde et al. (2016) investigated their models for further datasets and taking compound and constituent properties into account. Salehi et al. (2014) defined a cross-lingual distributional model that used translations into multiple languages and distributional similarities in the respective languages, to predict the compositionality for the two datasets from Reddy et al. (2011b) and Schulte im Walde et al. (2013). Cordeiro et al. (2019) provide the most recent investigation in a cross-linguistic study on the effects of corpus, modelling and composite parameters for English, French and Portuguese.

3. Data

3.1. Gold Standard of Noun Compounds

Our focus of interest is on English noun compounds, such as *butterfly, snowball* and *teaspoon* as well as *car park, zebra crossing* and *couch potato*,[1] where the grammatical head (in English, this is typically the rightmost constituent) is a noun. We are interested in the degrees of compositionality of noun compounds, i.e., the semantic relatedness between the meaning of a compound (e.g., *snowball*) and the meanings of its constituents (e.g., *snow* and *ball*).

As gold standard we used the dataset of English noun compounds created by Reddy et al. (2011b). Assuming that compounds whose constituents appeared either as their hypernyms or in their definitions tend to be compositional, Reddy et al. induced a candidate compound set with various degrees of compound–constituent relatedness from *WordNet* (Miller et al., 1990; Fellbaum, 1998) and *Wiktionary*. A random choice of 90 compounds that appeared with a corpus frequency > 50 in the *ukWaC* corpus (Baroni et al., 2009) constituted their gold-standard dataset and was annotated by compositionality ratings on the semantic contribution of the modifier to the compound meaning *(Word1)*, the semantic contribution of the head noun to the compound meaning *(Word2)*, and the compositionality of the compound as a whole *(Phrase)*. Table 1 shows some examples from their compounds and ratings.

Compound	Word1	Word2	Phrase
climate change	**4.90**±0.30	**4.83**±0.38	**4.97**±0.18
polo shirt	**1.73**±1.41	**5.00**±0.00	**3.37**±1.38
search engine	**4.62**±0.96	**2.25**±1.70	**3.32**±1.16
cheat sheet	**2.30**±1.59	**4.00**±0.83	**2.89**±1.11
gilt trip	**4.71**±0.59	**0.86**±0.94	**2.19**±1.16
night owl	**4.47**±0.88	**0.50**±0.82	**1.93**±1.27
crocodile tears	**0.19**±0.47	**3.79**±1.05	**1.25**±1.09
melting pot	**1.00**±1.15	**0.48**±0.63	**0.54**±0.63

Table 1: Examples of compounds and judgements on their compositionality (mean value and standard deviation, based on 30 annotators) from Reddy et al. (2011b), sorted by decreasing mean value of *Phrase*.

[1]Note that noun compounds in English may occur as closed compounds (without spaces), open compounds (with spaces) and hyphenated compounds, such as *butterfly, zebra crossing* and *long-term*, respectively. The benchmark dataset we used contains only open compounds.

3.2. Corpus and Co-Occurrence Vector Space

As corpus data for our vector-space variants we used one of the currently largest webcorpora for English: *ENCOW16*[2] containing ≈9.6 billion words (Schäfer and Bildhauer, 2012; Schäfer, 2015). We applied the *TreeTagger* for part-of-speech (pos) tagging and lemmatisation (Schmid, 1994), and we created frequency lists for all corpus lemmas and lemma-pos combinations.

As basis for our vector-space variants, we created a co-occurrence matrix for the gold-standard compounds and their constituents using a standard 10-word window (left+right) across the lemmatised ENCOW16. The window was applied within-sentence because the corpus is sentence-shuffled, such that going beyond sentence border is not meaningful. Since our target compounds are open compounds (with spaces) we pre-processed the corpus by joining all space-separated instances of the compounds in the corpus to represent a single token when running the window counts. The resulting target–context matrix contains 90 compound and 168 constituents targets (i.e., a total of 258 targets) as rows and 64,508 context dimensions across parts-of-speech as columns.

4. Experiments on Predicting Compositionality

4.1. Vector-Space Variants

Based on the general co-occurrence matrix described in the previous Section 3.2., we systematically created vector-space reductions across kinds, i.e., exploring part-of-speech-based reduction next to and also in combination with Principal Components Analysis (using Singular Value Decomposition) and word2vec embeddings. In the following, we describe our variants; Table 2 lists the variants accompanied by their dimensionality.

- **ALL**

  As baseline we used the whole co-occurrence matrix.

- **POS**

  We used subsets of the co-occurrence matrix with only context dimensions of specific parts-of-speech (specifying on nouns vs. verbs),[3] and from specific frequency ranges, as previously done by Schulte im Walde et al. (2013) in a similar way. Since nouns were generally more useful than verbs (see results below), we performed former fine-tuning just on the noun matrix by using only the 1,000/5,000/10,000/.../40,000 most frequent nouns from the corpus as context dimensions.[4]

[2]`http://corporafromtheweb.org/encow16/`

[3]In Schulte im Walde et al. (2013) we performed an elaborate investigation across parts-of-speech taking also adjectives and part-of-speech combinations into account.

[4]Note that the context dimensionalities of the matrices, i.e., the numbers of the context columns, are not necessarily equal to the number of the most frequent nouns, as not all of these nouns co-occurred as contexts with our targets.

- **PCA**

  We performed Principle Components Analysis (PCA) using Singular Value Decomposition (SVD) to reduce the dimensionality of the whole matrix and the matrices containing only noun dimensions.

  With this PCA-using-SVD method, our matrix M was first decomposed into three matrices: $M = U\Sigma W^T$ (i.e., performing Singular Value Decomposition). Then, when reducing the number of dimensions to k, we sliced U to the first k rows, Σ to the top-left $k \times k$ matrix, and W^T to the first k columns. Multiplying the three matrices provided a new matrix with less dimensions than previously in M.

- **WORD2VEC**

  We trained a standard word2vec two-layer neural network model (Mikolov et al., 2013) on the ENCOW16 corpus with window size 10 to obtain 300-dimensional word vectors for our compounds and constituents.

4.2. Prediction Functions

Relying on the vector-space variants, the *cosine* determined the distributional similarity between the compounds and their constituents, which was in turn used to predict the semantic relatedness between the compounds and their constituents, assuming that the stronger the distributional similarity (i.e., the higher the cosine values), the stronger the semantic relatedness and therefore the degree of compositionality.

Next to assessing the individual contributions of compound–modifier and compound–head relatedness, we applied the same functions as in Reddy et al. (2011b) to combine the compound–constituent cosine scores for predicting the degree of compositionality of the compounds, as also done in more general terms for in-depth investigations of phrase composite functions (Mitchell and Lapata, 2010; Coecke et al., 2011; Baroni et al., 2014a; Hermann, 2014):

WORD1 use only the compound–modifier cosine score

WORD2 use only the compound–head cosine score

ADD add the compound–modifier and compound–head cosine scores

MULT multiply the compound–modifier and compound–head cosine scores

COMB add the compound–modifier, the compound–head and the multiplication of both cosine scores

Given that each component within the functions might provide a different weight to the overall prediction, we used a linear regression model in order to predict the f function and to find the corresponding coefficients. After a 3-fold cross-validation with the human judgement, the best result is reported.

The vector space predictions were evaluated against the mean human ratings on the degree of compositionality, using the Spearman Rank-Order Correlation Coefficient ρ (Siegel and Castellan, 1988).

4.3. Taking Compound and Constituent Properties into Account

In order to zoom into specific strengths of individual vector space variants, we apply the variants to subsets of our compound targets according to the targets'

- degree of compositionality,
- compound frequency,
- modifier productivity, and
- head productivity.

For each of these conditions, we created three disjunctive subsets of the 90 compound targets with 30 targets each. The subsets contain the strongest, weakest and in-between targets as based on the respective condition, e.g., regarding the compound frequency condition we distinguish between high-frequency, mid-frequency and low-frequency compounds. The empirical information relies on a refinement of the Reddy et al. dataset by Schulte im Walde et al. (2016).

5. Experiment Results

Table 3 shows the overall results of predicting compositionality across our vector space variants and the prediction functions. The best-performing variants per kind of variation (as separated by horizontal lines) are in bold font.

We can see that the Word2Vec vector space outperforms all other variants with a correlation of $\rho = 0.689$. Obviously, this is not only a matter of dimensionality, as each reduction variant exhibits an individual behaviour regarding the optimal number of dimensions: The rather similar next-best results are reached with (i) using the most frequent corpus nouns NN-25000/NN-30000 (which effectively relies on 6,000–7,000 noun dimensions): $\rho = 0.663$; (ii) using only nouns (NN: all 52,285 of them): $\rho = 0.658$; and (iii) PCA on the noun-only matrix (NN-PCA), when using 2,000 dimensions: $\rho = 0.657$. Performing PCA on the whole matrix (All-PCA) is worse, reaching a maximum of $\rho = 0.616$ with 5,000 dimensions. A purely pos-based reduction for verbs-only reaches $\rho = 0.581$, in comparison to $\rho = 0.658$ for nouns-only, thus confirming the study by Schulte im Walde et al. (2013) in that nouns are more reliable than verbs in vector spaces for predicting compositionality. The baseline with using all context dimensions ($\rho = 0.630$) is worse in comparison to all reduced conditions other than running PCA on the whole matrix. Therefore, next to identifying a clear winner (Word2Vec) we can induce from our results that using only the most frequent noun dimensions is a reasonable alternative.

Regarding the prediction functions, ADD, MULT and COMB (with only marginal differences between them in most cases) are generally outperforming WORD1 and WORD2. So combining the relatedness information for compound–modifier and compound–head pairs is better for the prediction of the overall compounds' degree of compositionality than relying on just one or the other. Note that the predictions using the compound–head information (WORD2) are often strongly below the compound–modifier predictions (WORD1). For Word2Vec it is striking that MULT is doing a very poor job in comparison to all other functions.

	Variant	Dimensionality
All	all context words co-occurring with any of our targets	64,508
VV	all verbs	8,525
NN	all nouns	52,285
NN-1000	1,000 most frequent corpus nouns	374
NN-5000	5,000 most frequent ncorpus ouns	1,221
NN-10000	10,000 most frequent corpus nouns	2,392
NN-15000	15,000 most frequent corpus nouns	3,615
NN-20000	20,000 most frequent corpus nouns	4,762
NN-25000	25,000 most frequent corpus nouns	5,929
NN-30000	30,000 most frequent corpus nouns	6,970
NN-35000	35,000 most frequent corpus nouns	8,058
NN-40000	40,000 most frequent corpus nouns	9,114
All-PCA-100	PCA with 100 dimensions computed on whole matrix	100
All-PCA-500	PCA with 500 dimensions computed on whole matrix	500
All-PCA-1000	PCA with 1,000 dimensions computed on whole matrix	1,000
All-PCA-2000	PCA with 2,000 dimensions computed on whole matrix	2,000
All-PCA-5000	PCA with 5,000 dimensions computed on whole matrix	5,000
NN-PCA-100	PCA with 100 dimensions computed on noun matrix	100
NN-PCA-500	PCA with 500 dimensions computed on noun matrix	500
NN-PCA-1000	PCA with 1,000 dimensions computed on noun matrix	1,000
NN-PCA-2000	PCA with 2,000 dimensions computed on noun matrix	2,000
NN-PCA-5000	PCA with 5,000 dimensions computed on noun matrix	5,000
Word2Vec	word2vec two-layer neural network representation	300

Table 2: Vector-space variants and their dimensionalities.

Vector-Space Variants	WORD1	WORD2	ADD	MULT	COMB
All	0.583	0.444	**0.630**	0.626	**0.630**
Verbs (VV)	0.534	0.383	0.581	0.387	0.578
Nouns (NN)	0.634	0.433	**0.658**	**0.658**	0.655
NN-1000	0.436	0.324	0.482	0.452	0.483
NN-5000	0.614	0.377	0.630	0.592	0.630
NN-10000	0.618	0.397	0.638	0.632	0.637
NN-15000	0.631	0.429	0.653	0.648	0.652
NN-20000	0.637	0.435	0.661	0.659	0.658
NN-25000	0.640	0.438	**0.663**	**0.663**	0.662
NN-30000	0.641	0.437	0.662	**0.663**	0.662
NN-35000	0.633	0.433	0.656	0.652	0.653
NN-40000	0.635	0.432	0.657	0.659	0.656
All-PCA-100	0.456	0.321	0.527	0.487	0.504
All-PCA-500	0.510	0.357	0.584	0.573	0.577
All-PCA-1000	0.562	0.375	0.564	0.574	0.564
All-PCA-2000	0.554	0.432	0.601	0.604	0.609
All-PCA-5000	0.576	0.432	0.616	0.610	**0.616**
NN-PCA-100	0.536	0.320	0.620	0.587	0.613
NN-PCA-500	0.578	0.353	0.610	0.631	0.620
NN-PCA-1000	0.566	0.402	0.614	0.635	0.640
NN-PCA-2000	0.628	0.433	0.639	**0.657**	0.646
NN-PCA-5000	0.608	0.433	0.643	0.654	0.653
Word2Vec	0.602	0.435	0.680	0.145	**0.689**

Table 3: Results across vector-space variants and prediction functions.

In the following we now zoom into the results for specific subsets of the gold standard, distinguishing between low-/mid-/high-frequency compounds, compounds with low-/mid-/high-productivity modifiers vs. heads, and compounds with low-/mid-/high-compositionality phrases, modifiers and heads. In general, we observed that with training the regression on the whole dataset and testing it on the subsets we obtained the same results as with training the regression on the subsets.

The results on the subsets are shown in Tables 4–9. The best-performing variant per range is in bold font; in addition, the best-performing variant per reduction kind is highlighted by yellow background colour.

Results across Compound Frequency Ranges Zooming into the prediction results for high-, mid and low-frequency compounds (see Table 4), we first of all observe that Word2Vec by far outperforms the other reduction variants for high- and low-frequency compounds. In addition, the most striking differences in Table 4 in comparison to Table 3 are two-fold: On the one hand we can see that the prediction results for low-frequency compounds are much below those for mid-frequency and high-frequency compounds; only for Word2Vec this is not the case. On the other hand, the (rather low) best prediction results for the low-frequency compounds are achieved by WORD1 and WORD2 (again, this does not apply to Word2Vec but to all other kinds of reduction). Finally, in all but Word2Vec the prediction results for mid-frequency compounds are clearly above those for low- and high-frequency compounds.

Results distinguishing Modifier Productivity Ranges Zooming into the prediction results for compounds with high-, mid and low-productivity modifiers (see Table 5), we can see that differently to the previous cases here the nouns-only vector space provides the overall best results; this is the case for compounds with low-productivity modifiers. Overall, we can however not observe strong differences across reduction variants: several kinds of spaces are similarly successful across compound subsets. Interestingly, though, we observe much more variability in which prediction functions are best in predicting compositionality for compounds with low-, mid- and high-productive modifiers. Overall, WORD1, ADD, MULT and COMB take turns in being most successful, and there is no subset–function pairing that strikes as a particularly strong combination. So in sum, it is difficult to identify any tendencies of variants across modifier productivity subsets. This insight is in line with our previous work (Schulte im Walde et al., 2016) which also demonstrated that empirical modifier properties do not have a consistent effect on the quality of predicting compound compositionality.

Results distinguishing Head Productivity Ranges In contrast, zooming into the prediction results for compounds with high-, mid and low-productivity heads (see Table 6), we do observe patterns for compound subsets. In all chosen space variants, the prediction is best for compounds with mid-productivity heads, second-best for those with high-productivity heads and worst for those with low-productivity heads. This is surprising on the one hand, given that mid-range ratings typically show higher standard deviations and less agreement across human raters (Pollock, 2018), so one might consider their degrees of compositionality more difficult to distinguish than others. On the other hand, compounds with low-productivity heads are supposedly more influenced by sparse data in the vectors, and this does not seem to change in dimensionality-reduced vector spaces.
Comparing vector variants and prediction functions, Word2Vec is again the best option but the noun-based variants NN and NN-PCA are similarly successful. ADD, MULT and COMB are mostly the best functions, but in individual low-productivity cases WORD1 and WORD2 are best.

Results distinguishing Compositionality Ranges Finally, Tables 7–9 zoom into prediction results across degrees of compositionality, regarding the compound phrase as a whole (Table 7), the compound–modifier relation (Table 8), and the compound–head relation (Table 9). For predictions across degrees of phrase compositionality (Table 7), Word2Vec is the clear winner for high- and low-compositional compounds, and for mid-compositional compounds both NN-PCA and Word2Vec clearly outperform the other functions. For high-compositional compounds, WORD1 is the best prediction function, so modifiers seem to determine the prediction in high-compositional cases. Otherwise ADD, MULT and COMB represent the best functions, as before.
For compounds with varying modifier or head compositionality the picture is more diverse. What is most interesting here is that for compounds with low-compositional modifiers (Table 8) WORD2 represents the best prediction function, while in most cases in Table 9, WORD1 represents the best prediction function. We interpret this behaviour as follows: For compounds with low-compositional modifiers the semantic relatedness compound–modifier is low, and here the strength of semantic relatedness compound–head (which is effectively WORD2) correlates with the degree of compositionality of the phrase. Thus, in cases with low compound–modifier relatedness the degree of compositionality of the compound phrase and the compound–head pair are similar in their ranks across compounds. When investigating compounds with varying degrees of head compositionality this effect even applies across compound–head ranges of compositionality, i.e., the strength of semantic relatedness compound–modifier (which is effectively WORD1) correlates with the degree of compositionality of the phrase, so the degree of compositionality of the compound phrase and the compound–modifier pair are similar in their ranks within all three ranges.

6. Summary and Conclusion

This study provided a systematic evaluation of vector-space reductions across kinds, i.e., exploring part-of-speech-based reduction, Principal Components Analysis using Singular Value Decomposition, and word2vec embeddings.

Vector-Space Variants	WORD1	WORD2	ADD	MULT	COMB
High-frequency compounds					
All	0.362	0.153	0.393	0.409	0.399
NN	0.365	0.152	0.366	0.337	0.372
All-PCA-1000	0.359	0.153	0.366	0.396	0.367
NN-PCA-2000	0.365	0.152	0.372	0.337	0.447
Word2Vec	0.616	0.378	0.642	0.551	**0.678**
Mid-frequency compounds					
All	0.365	0.256	0.470	0.451	0.441
NN	0.578	0.298	0.606	0.570	0.607
All-PCA-5000	0.365	0.256	0.461	0.451	0.484
NN-PCA-2000	0.578	0.298	**0.616**	0.570	0.613
Word2Vec	0.518	0.440	0.585	0.565	0.577
Low-frequency compounds					
All	0.268	0.132	0.178	0.179	0.205
NN	0.330	0.208	0.268	0.218	0.208
All-PCA-5000	0.196	0.040	0.154	0.087	0.118
NN-PCA-2000	0.233	0.262	0.127	0.250	0.174
Word2Vec	0.314	0.140	**0.352**	0.301	0.334

Table 4: Results distinguishing compound frequency ranges.

Vector-Space Variants	WORD1	WORD2	ADD	MULT	COMB
High-productivity modifiers					
All	0.428	0.411	0.569	0.594	0.534
NN	0.407	0.370	0.494	0.543	0.538
All-PCA-5000	0.428	0.415	0.599	0.594	0.600
NN-PCA-2000	0.407	0.370	0.522	0.543	0.526
Word2Vec	**0.631**	0.221	0.568	0.473	0.598
Mid-productivity modifiers					
All	0.568	0.447	0.580	0.576	0.619
NN	0.407	0.370	0.494	0.543	0.538
All-PCA-5000	0.532	0.378	0.554	0.513	0.530
NN-PCA-2000	0.603	0.442	0.609	0.594	**0.625**
Word2Vec	0.566	0.243	0.571	0.385	0.563
Low-productivity modifiers					
All	0.394	0.134	0.321	0.382	0.388
NN	0.632	0.478	**0.653**	0.628	0.634
All-PCA-5000	0.414	0.134	0.294	0.361	0.343
NN-PCA-2000	0.636	0.173	0.217	0.554	0.445
Word2Vec	0.385	0.514	0.485	0.584	0.471

Table 5: Results distinguishing modifier productivity ranges.

Relying on the gold standard of English noun compounds by Reddy et al. (2011b), our vector-space variant experiments identified word2vec with 300 dimensions as the clear winner. Similarly good and stable predictions have been achieved when using a large subset of context nouns (in our case relying on the ca. 25,000–30,000 most frequent out of a total of ca. 50,000 noun types), with or without any further PCA reduction.

Zooming into prediction functions and compound and constituent properties, we further demonstrated that –while the overall best predictions are performed with function combination (addition, multiplication, combination of both)– the picture varies strongly across subsets representing different ranges of compositionality, frequency and productivity:

1. Predictions for low-frequency compounds are much worse, and predictions for mid-frequency compounds are much better than on average.

2. There are no obvious tendencies across modifier productivity ranges, but for head productivity ranges we observe very high prediction results for mid-productivity, very low prediction results for low-productivity, and medium prediction results for high-productivity subsets.

Vector-Space Variants	WORD1	WORD2	ADD	MULT	COMB
High-productivity heads					
All	0.524	0.313	0.540	0.555	0.545
NN	0.547	0.433	0.593	0.648	0.615
All-PCA-5000	0.552	0.322	0.545	0.559	0.531
NN-PCA-2000	0.433	0.437	0.611	0.677	0.608
Word2Vec	0.643	0.420	0.686	0.667	**0.701**
Mid-productivity heads					
All	0.708	0.449	0.784	0.698	0.758
NN	0.727	0.474	**0.840**	0.789	0.823
All-PCA-5000	0.708	0.449	0.746	0.698	0.739
NN-PCA-2000	0.727	0.474	0.827	0.789	0.821
Word2Vec	0.639	0.532	0.801	0.718	0.791
Low-productivity heads					
All	0.155	0.245	0.181	0.191	0.151
NN	0.337	0.227	0.292	0.259	0.225
All-PCA-5000	0.124	0.227	0.175	0.152	0.199
NN-PCA-2000	0.293	0.204	0.249	0.316	0.302
Word2Vec	0.370	0.425	0.472	0.468	**0.474**

Table 6: Results distinguishing head productivity ranges.

Vector-Space Variants	WORD1	WORD2	ADD	MULT	COMB
High-compositional compounds					
All	0.240	0.008	0.206	0.230	0.216
NN	0.274	0.118	0.290	0.293	0.287
All-PCA-5000	0.362	0.153	0.404	0.409	0.393
NN-PCA-2000	0.365	0.152	0.375	0.337	0.361
Word2Vec	**0.631**	0.221	0.568	0.473	0.598
Mid-compositional compounds					
All	0.130	0.253	0.223	0.256	0.210
NN	0.230	0.245	0.247	0.290	0.287
All-PCA-5000	0.196	0.040	0.094	0.087	0.099
NN-PCA-2000	0.578	0.298	0.610	0.570	**0.614**
Word2Vec	0.566	0.243	0.571	0.385	0.563
Low-compositional compounds					
All	0.420	0.407	0.514	0.489	0.525
NN	0.426	0.287	0.469	0.423	0.452
All-PCA-5000	0.196	0.040	0.094	0.087	0.099
NN-PCA-2000	0.231	0.242	0.148	0.259	0.190
Word2Vec	0.385	0.514	0.485	**0.584**	0.471

Table 7: Results distinguishing compound compositionality ranges.

Vector-Space Variants	WORD1	WORD2	ADD	MULT	COMB
High-compositional modifiers					
All	0.285	0.476	0.471	0.498	0.420
NN	0.408	0.551	0.520	0.620	0.574
All-PCA-5000	0.323	0.516	0.472	0.535	0.343
NN-PCA-2000	0.430	0.444	0.611	0.649	0.569
Word2Vec	0.606	0.350	0.625	0.560	**0.650**
Mid-compositional modifiers					
All	0.306	0.534	0.560	0.535	0.471
NN	0.500	0.646	0.590	0.662	0.653
All-PCA-5000	0.254	0.550	0.584	0.532	0.601
NN-PCA-2000	0.402	0.654	0.662	**0.681**	0.660
Word2Vec	0.345	0.556	0.536	0.414	0.585
Low-compositional modifiers					
All	0.302	**0.434**	0.372	0.403	0.370
NN	0.252	0.343	0.245	0.256	0.284
All-PCA-5000	0.224	**0.434**	0.398	0.331	0.430
NN-PCA-2000	0.280	0.281	0.271	0.202	0.363
Word2Vec	0.214	0.417	0.372	0.350	0.373

Table 8: Results distinguishing modifier compositionality ranges.

Vector-Space Variants	WORD1	WORD2	ADD	MULT	COMB
High-compositional heads					
All	0.777	0.171	0.700	0.724	0.735
NN	0.752	0.199	0.735	0.736	0.713
All-PCA-5000	**0.779**	0.162	0.722	0.723	0.734
NN-PCA-2000	0.753	0.172	0.722	0.737	0.744
Word2Vec	0.761	0.064	0.644	0.524	0.678
Mid-compositional heads					
All	0.476	-0.033	0.371	0.345	0.427
NN	0.589	-0.025	0.512	0.340	0.437
All-PCA-5000	0.518	0.011	0.467	0.327	0.519
NN-PCA-2000	0.578	0.071	0.585	0.457	0.538
Word2Vec	0.525	0.338	**0.592**	0.498	0.583
Low-compositional heads					
All	0.361	-0.014	0.267	0.249	0.269
NN	**0.496**	-0.080	0.407	0.308	0.339
All-PCA-5000	0.253	-0.134	0.011	0.108	0.011
NN-PCA-2000	0.468	0.046	0.272	0.330	0.201
Word2Vec	0.160	0.071	0.288	0.191	0.279

Table 9: Results distinguishing head compositionality ranges.

3. For compounds with low compound–modifier relatedness the compound–head relatedness can be used for predicting the overall compound phrase compositionality; even stronger, the compound–modifier relatedness can be used for predicting the overall compound phrase compositionality for compounds across compound–head relatedness ranges.

Many of these insights correpond to those in Schulte im Walde et al. (2016) and once more emphasise the importance of balancing target properties in gold standards. Especially the latter results call for further work on other datasets and across languages.

7. Bibliographical References

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, M., Bernardi, R., and Zamparelli, R. (2014a). Frege in Space: A Program for Compositional Distributional Semantics. *Linguistic Issues in Language Technologies*, 9(6):5–110.

Baroni, M., Dinu, G., and Kruszewski, G. (2014b). Don't count, predict! A Systematic Comparison of Context-counting and Context-predicting Semantic Vectors. In

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 238–247, Baltimore, MD.

Bullinaria, J. A. and Levy, J. P. (2012). Extracting Semantic Representations from Word Co-Occurrence Statistics: Stop-Lists, Stemming, and SVD. *Behavior Research Methods*, 44:890–907.

Cap, F., Nirmal, M., Weller, M., and Schulte im Walde, S. (2015). How to Account for Idiomatic German Support Verb Constructions in Statistical Machine Translation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 19–28, Denver, Colorado, USA.

Cholakov, K. and Kordoni, V. (2014). Better Statistical Machine Translation through Linguistic Treatment of Phrasal Verbs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 196–201, Doha, Qatar.

Coecke, B., Sadrzadeh, M., and Clark, S. (2011). Mathematical Foundations for a Compositional Distributional Model of Meaning. *Linguistic Analysis*, 36(1-4):345–384.

Cordeiro, S., Villavicencio, A., Idiart, M., and Ramisch, C. (2019). Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 45(1):1–57.

Christiane Fellbaum, editor. (1998). *WordNet – An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA, USA.

Hermann, K. M. (2014). *Distributed Representations for Compositional Semantics*. Ph.D. thesis, University of Oxford.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of Computational Linguistics*, 3:211–225.

Mikolov, T., tau Yih, W., and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, GA, USA.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to Wordnet: An Online Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

Mitchell, J. and Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, 34:1388–1429.

Pollock, L. (2018). Statistical and Methodological Problems with Concreteness and other Semantic Variables: A List Memory Experiment Case Study. *Behavior Research Methods*, 50:1198–1216.

Reddy, S., Klapaftis, I. P., McCarthy, D., and Manandhar, S. (2011a). Dynamic and Static Prototype Vectors for Semantic Composition. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 705–713, Chiang Mai, Thailand.

Reddy, S., McCarthy, D., and Manandhar, S. (2011b). An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand.

Salehi, B. and Cook, P. (2013). Predicting the Compositionality of Multiword Expressions Using Translations in Multiple Languages. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pages 266–275, Atlanta, GA, USA.

Salehi, B., Cook, P., and Baldwin, T. (2014). Using Distributional Similarity of Multi-way Translations to Predict Multiword Expression Compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden.

Salehi, B., Cook, P., and Baldwin, T. (2015a). A Word Embedding Approach to Predicting the Compositionality of Multiword Expressions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies*, pages 977–983, Denver, Colorado, USA.

Salehi, B., Mathur, N., Cook, P., and Baldwin, T. (2015b). The Impact of Multiword Expression Compositionality on Machine Translation Evaluation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 54–59, Denver, Colorado, USA.

Schäfer, R. and Bildhauer, F. (2012). Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey.

Schäfer, R. (2015). Processing and Querying Large Web Corpora with the COW14 Architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28–34, Mannheim, Germany.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging using Decision Trees. In *Proceedings of the 1st International Conference on New Methods in Language Processing*.

Schulte im Walde, S., Müller, S., and Roller, S. (2013). Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pages 255–265, Atlanta, GA, USA.

Schulte im Walde, S., Hätty, A., and Bott, S. (2016). The Role of Modifier and Head Properties in Predicting the Compositionality of English and German Noun-Noun Compounds: A Vector-Space Perspective. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, pages 148–158, Berlin, Germany.

Siegel, S. and Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Boston, MA, USA.

Weller, M., Cap, F., Müller, S., Schulte im Walde, S., and Fraser, A. (2014). Distinguishing Degrees of Compositionality in Compound Splitting for Statistical Machine Translation. In *Proceedings of the 1st Workshop on Computational Approaches to Compound Analysis*, pages 81–90, Dublin, Ireland.