Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space

Marco Baroni and **Roberto Zamparelli**
Center for Mind/Brain Sciences, University of Trento
Rovereto (TN), Italy
{marco.baroni,roberto.zamparelli}@unitn.it

Abstract

We propose an approach to adjective-noun composition (AN) for corpus-based distributional semantics that, building on insights from theoretical linguistics, represents nouns as vectors and adjectives as data-induced (linear) functions (encoded as matrices) over nominal vectors. Our model significantly outperforms the rivals on the task of reconstructing AN vectors not seen in training. A small post-hoc analysis further suggests that, when the model-generated AN vector is not similar to the corpus-observed AN vector, this is due to anomalies in the latter. We show moreover that our approach provides two novel ways to represent adjective meanings, alternative to its representation via corpus-based co-occurrence vectors, both outperforming the latter in an adjective clustering task.

1 Introduction

An influential approach for representing the meaning of a word in NLP is to treat it as a vector that codes the pattern of co-occurrence of that word with other expressions in a large corpus of language (Sahlgren, 2006; Turney and Pantel, 2010). This approach to semantics (sometimes called *distributional semantics*) naturally captures word clustering, scales well to large lexicons and doesn't require words to be manually disambiguated (Schütze, 1997). However, until recently it has been limited to the level of content words (nouns, adjectives, verbs), and it hasn't tackled in a general way *compositionality* (Frege, 1892; Partee, 2004), that crucial property of natural language which allows speakers to derive the meaning of a complex linguistic constituent from the meaning of its immediate syntactic subconstituents.

Formal semantics (FS), the research program stemming from Montague (1970b; 1973), has opposite strengths and weaknesses. Its core semantic notion is the sentence, not the word; at the lexical level, it focuses on the meaning of function words; one of its main goals is to formulate recursive compositional rules that derive the quantificational properties of complex sentences and their antecedent-pronoun dependencies.

Given its focus on quantification, FS treats the meanings of nouns and verbs as pure *extensions*: nouns and (intransitive) verbs are properties, and thus denote sets of individuals. Adjectives are also often assumed to denote properties: in this view red_{adj} would be the set of 'entities which are red', $plastic_{adj}$, the set of 'objects made of plastic', and so forth. In the simplest case, the meaning of an attributive adjective-noun (AN) constituent can be obtained as the intersection of the adjective and noun extensions A∩N:

$$[\text{ red car }] = \{\ldots red\ objects\ldots\} \cap \{\ldots cars\ldots\}$$

However, the intersective method of combination is well-known to fail in many cases (Kamp, 1975; Montague, 1970a; Siegel, 1976): for instance, a *fake gun* is not a *gun*. Even for *red*, the manner in which the color combines with a noun will be different in *red Ferrari* (the outside), *red watermelon* (the inside), *red traffic light* (the signal). These problems have prompted a more flexible FS representation for attributive adjectives — functions from the meaning of a noun onto the meaning of a modified noun (Montague, 1970a). This mapping could now be sensitive to the particular noun the adjective receives, and it does not need to return a subset of the

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1183–1193,
MIT, Massachusetts, USA, 9-11 October 2010.

original noun denotation (as in the case of *fake N*). However, FS has nothing to say on how these functions should be constructed.

In the last few years there have been attempts to build compositional models that use distributional semantic representations as inputs (see Section 2 below), most of them focusing on the combination of a verb and its arguments. This paper addresses instead the combination of nouns and attributive adjectives. This case was chosen as an interesting testbed because it has the property of recursivity (it applies in *black dog*, but also in *large black dog*, etc.), and because very frequent adjectives such as *different* are at the border between content and function words. Following the insight of FS, we treat attributive adjectives as *functions* over noun meanings; however, noun meanings are vectors, not sets, and the functions are learnt from corpus-based noun-AN vector pairs.

Original contribution We propose and evaluate a new method to derive distributional representations for ANs, where an adjective is a linear function from a vector (the noun representation) to another vector (the AN representation). The linear map for a specific adjective is learnt, using linear regression, from pairs of noun and AN vectors extracted from a corpus.

Outline Distributional approaches to compositionality are shortly reviewed in Section 2. In Section 3, we introduce our proposal. The experimental setting is described in Section 4. Section 5 provides some empirical justification for using corpus-harvested AN vectors as the target of our function learning and evaluation benchmark. In Section 6, we show that our model outperforms other approaches at the task of approximating such vectors for unseen ANs. In Section 7, we discuss how adjectival meaning can be represented in our model and evaluate this representation in an adjective clustering task. Section 8 concludes by sketching directions for further work.

2 Related work

The literature on compositionality in vector-based semantics encompasses various related topics, some of them not of direct interest here, such as how to encode word order information in context vectors (Jones and Mewhort, 2007; Sahlgren et al., 2008) or sophisticated composition methods based on tensor products, quantum logic, etc., that have not yet been empirically tested on large-scale corpus-based semantic space tasks (Clark and Pulman, 2007; Rudolph and Giesbrecht, 2010; Smolensky, 1990; Widdows, 2008). Closer to our current purposes is the general framework for vector composition proposed by Mitchell and Lapata (2008), subsuming various earlier proposals. Given two vectors $\mathbf{u}$ and $\mathbf{v}$, they identify two general classes of composition models, (linear) additive models:

$$\mathbf{p} = \mathbf{Au} + \mathbf{Bv} \quad (1)$$

where $\mathbf{A}$ and $\mathbf{B}$ are weight matrices, and multiplicative models:

$$\mathbf{p} = \mathbf{Cuv}$$

where $\mathbf{C}$ is a weight tensor projecting the $\mathbf{uv}$ tensor product onto the space of $\mathbf{p}$. Mitchell and Lapata derive two simplified models from these general forms. Their simplified additive model $\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v}$ was a common approach to composition in the earlier literature, typically with the scalar weights set to 1 or to normalizing constants (Foltz et al., 1998; Kintsch, 2001; Landauer and Dumais, 1997). Mitchell and Lapata also consider a constrained version of the multiplicative approach that reduces to component-wise multiplication, where the i-th component of the composed vector is given by: $p_i = u_i v_i$. The simplified additive model produces a sort of (statistical) union of features, whereas component-wise multiplication has an intersective effect. They also evaluate a weighted combination of the simplified additive and multiplicative functions. The best results on the task of paraphrasing noun-verb combinations with ambiguous verbs (*sales slump* is more like *declining* than *slouching*) are obtained using the multiplicative approach, and by weighted combination of addition and multiplication (we do not test model combinations in our current experiments). The multiplicative approach also performs best (but only by a small margin) in a later application to language modeling (Mitchell and Lapata, 2009). Erk and Padó (2008; 2009) adopt the same formalism but focus on the nature of input vectors, suggesting that when a verb is composed with a noun, the noun component is given by an average of verbs that the noun is typically object of (along similar lines,

Kintsch (2001) also focused on composite input vectors, within an additive framework). Again, the multiplicative model works best in Erk and Padó's experiments.

The above-mentioned researchers do not exploit corpus evidence about the $\mathbf{p}$ vectors that result from composition, despite the fact that it is straightforward (at least for short constructions) to extract direct distributional evidence about the composite items from the corpus (just collect co-occurrence information for the composite item from windows around the contexts in which it occurs). The main innovation of Guevara (2010), who focuses on adjective-noun combinations (AN), is to use the co-occurrence vectors of observed ANs to train a supervised composition model (we became aware of Guevara's approach after we had developed our own model, that also exploits observed ANs for training). Guevara adopts the full additive composition form from Equation (1) and he estimates the $\mathbf{A}$ and $\mathbf{B}$ weights using partial least squares regression. The training data are pairs of adjective-noun vector concatenations, as input, and corpus-derived AN vectors, as output. Guevara compares his model to the simplified additive and multiplicative models of Mitchell and Lapata. Observed ANs are nearer, in the space of observed and predicted test set ANs, to the ANs generated by his model than to those from the alternative approaches. The additive model, on the other hand, is best in terms of shared neighbor count between observed and predicted ANs.

In our empirical tests, we compare our approach to the simplified additive and multiplicative models of Mitchell and Lapata (the former with normalization constants as scalar weights) as well as to Guevara's approach.

3 Adjectives as linear maps

As discussed in the introduction, we will take adjectives in attributive position to be functions from one noun meaning to another. To start simple, we assume here that adjectives in the attributive position (AN) are linear functions from n-dimensional (noun) vectors onto n-dimensional vectors, an operation that can be expressed as multiplication of the input noun column vector by a $n \times n$ matrix, that is our representation for the adjective (in the language of linear algebra, an adjective is an endomorphic linear map in noun space). In the framework of Mitchell and Lapata, our approach derives from the additive form in Equation (1) with the matrix multiplying the adjective vector (say, $\mathbf{A}$) set to $\mathbf{0}$:

$$\mathbf{p} = \mathbf{B}\mathbf{v}$$

where $\mathbf{p}$ is the observed AN vector, $\mathbf{B}$ the weight matrix representing the adjective at hand, and $\mathbf{v}$ a noun vector. In our approach, the weight matrix $\mathbf{B}$ is specific to a single adjective – as we will see in Section 7 below, it is our representation of the meaning of the adjective.

Like Guevara, we estimate the values in the weight matrix by partial least squares regression. In our case, the independent variables for the regression equations are the dimensions of the corpus-based vectors of the component nouns, whereas the AN vectors provide the dependent variables. Unlike Guevara, (i) we train separate models for each adjective (we learn adjective-specific functions, whereas Guevara learns a generic "AN-slot" function) and, consequently, (ii) corpus-harvested adjective vectors play no role for us (their values would be constant across the training input vectors).

A few considerations are in order. First, although we use a supervised learning method (least squares regression), we do not need hand-annotated data, since the target AN vectors are automatically collected from the corpus just like vectors for single words are. Thus, there is no extra "external knowledge" cost with respect to unsupervised approaches. Second, our approach rests on the assumption that the corpus-derived AN vectors are interesting objects that should constitute the target of what a composition process tries to approximate. We provide preliminary empirical support for this assumption in Section 5 below. Third, we have some reasonable hope that our functions can capture to a certain extent the polysemous nature of adjectives: we could learn, for example, a *green* matrix with large positive weights mapping from noun features that pertain to concrete objects to color dimensions of the output vector (*green chair*), as well as large positive weights from features characterizing certain classes of abstract concepts to political/social dimensions in the output (*green initiative*). Somewhat optimistically, we hope that *chair* will have near-0 values

on the relevant abstract dimensions, like *initiative* on the concrete features, and thus the weights will not interfere. We do not evaluate this claim specifically, but our quantitative evaluation in Section 6 shows that our approach does best with high frequency, highly ambiguous adjectives. Fourth, the approach is naturally syntax-sensitive, since we train it on observed data for a specific syntactic position: we would train separate linear models for, say, the same adjective in attributive (AN) and predicative (N is A) position. As a matter of fact, the current model is too syntax-sensitive and does not capture similarities across different constructions. Finally, although adjective representations are not directly harvested from corpora, we can still meaningfully compare adjectives to each other or other words by using their estimated matrix, or an average vector for the ANs that contain them: both options are tested in Section 7 below.

4 Experimental setup

4.1 Corpus

We built a large corpus by concatenating the Web-derived ukWaC corpus (`http://wacky.sslmit.unibo.it/`), a mid-2009 dump of the English Wikipedia (`http://en.wikipedia.org`) and the British National Corpus (`http://www.natcorp.ox.ac.uk/`). This **concatenated corpus**, tokenized, POS-tagged and lemmatized with the TreeTagger (Schmid, 1995), contains about 2.83 billion tokens (excluding punctuation, digits, etc.). The ukWaC and Wikipedia sections can be freely downloaded, with full annotation, from the ukWaC site.

We performed some of the list extraction and checking operations we are about to describe on a more manageable data-set obtained by selecting the first 100M tokens of ukWaC; we refer to this subset as the **sample corpus** below.

4.2 Vocabulary

We could in principle limit ourselves to collecting vectors for the ANs to be analyzed (the *AN test set*) and their components. However, to make the analysis more challenging and interesting, we populate the semantic space where we will look at the behaviour of the ANs with a large number of adjectives and nouns, as well as further ANs not in the test set. We refer to the overall list of items we build semantic vectors for as the *extended vocabulary*. We use a subset of the extended vocabulary containing only nouns and adjectives (the *core vocabulary*) for feature selection and dimensionality reduction, so that we do not implicitly bias the structure of the semantic space by our choice of ANs.

To construct the **AN test set**, we first selected 36 adjectives across various classes: size (*big, great, huge, large, major, small, little*), denominal (*American, European, national, mental, historical, electronic*), colors (*white, black, red, green*) positive evaluation (*nice, excellent, important, appropriate*), temporal (*old, recent, new, young, current*), modal (*necessary, possible*), plus some common abstract antonymous pairs (*difficult, easy, good, bad, special, general, different, common*). We were careful to include intersective cases such as *electronic* as well as non-intersective adjectives that are almost function words (the modals, *different*, etc.). We extracted all nouns that occurred at least 300 times in post-adjectival position in the sample corpus, excluding some extremely frequent temporal and measure expressions such as *time* and *range*, for a total of 1,420 distinct nouns. By crossing the selected adjectives and nouns, we constructed a test set containing 26,440 ANs, all attested in the sample corpus (734 ANs per adjective on average, ranging from 1,337 for *new* to 202 for *mental*).

The **core vocabulary** contains the top 8K most frequent noun lemmas and top 4K adjective lemmas from the concatenated corpus (excluding the top 50 most frequent nouns and adjectives). The **extended vocabulary** contains this core plus (i) the 26,440 test ANs, (ii) the 16 adjectives and 43 nouns that are components of these ANs and that are not in the core set, and (iii) 2,500 more ANs randomly sampled from those that are attested in the sample corpus, have a noun from the same list used for the test set ANs, and an adjective that occurred at least 5K times in the sample corpus. In total, the extended vocabulary contains 40,999 entries: 8,043 nouns, 4,016 adjectives and 28,940 ANs.

4.3 Semantic space construction

Full co-occurrence matrix The 10K lemmas (nouns, adjectives or verbs) that co-occur with

the largest number of items in the core vocabulary constitute the dimensions (columns) of our co-occurrence matrix. Using the concatenated corpus, we extract sentence-internal co-occurrence counts of all the items in the extended vocabulary with the 10K dimension words. We then transform the raw counts into Local Mutual Information (LMI) scores (LMI is an association measure that closely approximates the Log-Likelihood Ratio, see Evert (2005)).

Dimensionality reduction Since, for each test set adjective, we need to estimate a regression model for each dimension, we want a compact space with relatively few, dense dimensions. A natural way to do this is to apply the Singular Value Decomposition (SVD) to the co-occurrence matrix, and represent the items of interest with their coordinates in the space spanned by the first n right singular vectors. Applying SVD is independently justified because, besides mitigating the dimensionality problem, it often improves the quality of the semantic space (Landauer and Dumais, 1997; Rapp, 2003; Schütze, 1997). To avoid bias in favour of dimensions that capture variance in the test set ANs, we applied SVD to the core vocabulary subset of the co-occurrence matrix (containing only adjective and noun rows). The core 12K×10K matrix was reduced using SVD to a 12K×300 matrix. The other row vectors of the full co-occurrence matrix (including the ANs) were projected onto the same reduced space by multiplying them by a matrix containing the first n right singular vectors as columns. Merging the items used to compute the SVD and those projected onto the resulting space, we obtain a 40,999×300 matrix representing 8,043 nouns, 4,016 adjectives and 28,940 ANs. This reduced matrix constitutes a realistically sized semantic space, that also contains many items that are not part of our test set, but will be potential neighbors of the observed and predicted test ANs in the experiments to follow. The quality of the SVD reduction itself was independently validated on a standard similarity judgment data-set (Rubenstein and Goodenough, 1965), obtaining similar (and state-of-the-art-range) Pearson correlations of vector cosines and human judgments in both the original ($r = .70$) and reduced ($r = .72$) spaces.

There are several parameters involved in constructing a semantic space (choice of full and reduced dimensions, co-occurrence span, weighting method). Since our current focus is on alternative composition methods evaluated on a shared semantic space, exploring parameters pertaining to the construction of the semantic space is not one of our priorities, although we cannot of course exclude that the nature of the underlying semantic space affects different composition methods differently.

4.4 Composition methods

In the proposed **adjective-specific linear map** (*alm*) method, an AN is generated by multiplying an adjective weight matrix with a noun (column) vector. The j weights in the i-th row of the matrix are the coefficients of a linear regression predicting the values of the i-th dimension of the AN vector as a linear combination of the j dimensions of the component noun. The linear regression coefficients are estimated separately for each of the 36 tested adjectives from the corpus-observed noun-AN pairs containing that adjective (observed adjective vectors are not used). Since we are working in the 300-dimensional right singular vector space, for each adjective we have 300 regression problems with 300 independent variables, and the training data (the noun-AN pairs available for each test set adjective) range from about 200 to more than 1K items. We estimate the coefficients using (multivariate) partial least squares regression (PLSR) as implemented in the R `pls` package (Mevik and Wehrens, 2007). With respect to standard least squares estimation, this technique is more robust against over-training by effectively using a smaller number of orthogonal "latent" variables as predictors (Hastie et al., 2009, Section 3.4), and it exploits the multivariate nature of the problem (different regressions for each AN vector dimension to be predicted) when determining the latent dimensions. The number of latent variables to be used in the core regression are a free parameter of PLSR. For efficiency reasons, we did not optimize it. We picked instead 50 latent variables, by the rule-of-thumb reasoning that for any adjective we can use at least 200 noun-AN pairs for training, and the independent-variable-to-training-item ratio will thus never be above 1/4. We adopt a leave-one-out training regime, so that each target AN is generated by an adjective matrix that was estimated from all the

other ANs with the same adjective, minus the target.

We use PLSR with 50 latent variables also for our re-implementation of Guevara's (2010) **single linear map** (*slm*) approach, in which a single regression matrix is estimated for all ANs across adjectives. The training data in this case are given by the concatenation of the observed adjective and noun vectors (600 independent variables) coupled with the corresponding AN vectors (300 dependent variables). For each target AN, we randomly sample 2,000 other adjective-noun-AN tuples for training (with larger training sets we run into memory problems), and use the resulting coefficient matrix to generate the AN vector from the concatenated target adjective and noun vectors.

Additive AN vectors (*add* method) are obtained by summing the corresponding adjective and noun vectors after normalizing them (non-normalized addition was also tried, but it did not work nearly as well as the normalized variant). **Multiplicative** vectors (*mult* method) were obtained by component-wise multiplication of the adjective and noun vectors (normalization does not matter here since it amounts to multiplying the composite vector by a scalar, and the cosine similarity measure we use is scale-invariant). Finally, the *adj* and *noun* baselines use the **adjective** and **noun** vectors, respectively, as surrogates of the AN vector.

For the *add*, *mult*, *adj* and *noun* methods, we ran the tests of Section 6 not only in the SVD-reduced space, but also in the original 10K-dimensional co-occurrence space. Only the *mult* method achieved better performance in the original space. We conjecture that this is because the SVD dimensions can have negative values, leading to counter-intuitive results with component-wise multiplication (multiplying large opposite-sign values results in large negative values). We tried to alleviate this problem by assigning a 0 to composite dimensions where the two input vectors had different signs. The resulting performance was better but still below that of *mult* in original space. Thus, in Section 6 we report *mult* results from the full co-occurrence matrix; reduced space results for all other methods.

5 Study 1: ANs in semantic space

The actual distribution of ANs in the corpus, as recorded by their co-occurrence vectors, is fundamental to what we are doing. Our method relies on the hypothesis that the semantics of AN composition does not depend on the independent distribution of adjectives themselves, but on how adjectives transform the distribution of nouns, as evidenced by observed pairs of noun-AN vectors. Moreover, coherently with this view, our evaluation below will be based on how closely the models approximate the observed vectors of unseen ANs.

That our goal in modeling composition should be to approximate the vectors of observed ANs is in a sense almost trivial. Whether we synthesize an AN for generation or decoding purposes, we would want the synthetic AN to look as much as possible like a real AN in its natural usage contexts, and co-occurrence vectors of observed ANs are a summary of their usage in actual linguistic contexts. However, it might be the case that the specific resources we used for our vector construction procedure are not appropriate, so that the specific observed AN vectors we extract are not reliable (e.g., they are so sparse in the original space as to be uninformative, or they are strictly tied to the domains of the input corpora). We provide here some preliminary qualitative evidence that this is in general not the case, by tapping into our own intuitions on where ANs should be located in semantic space, and thus on how sensible their neighbors are.

First, we computed centroids from normalized SVD space vectors of all the ANs that share the same adjective (e.g., the normalized vectors of *American adult*, *American menu*, etc., summed to construct the *American N* centroid). We looked at the nearest neighbors of these centroids in semantic space among the 41K items (adjectives, nouns and ANs) in our extended vocabulary (here and in all experiments below, similarity is quantified by the cosine of the angle between two vectors). As illustrated for a random sample of 9 centroids in Table 1 (but applying to the remaining 27 adjectives as well), centroids are positioned in intuitively reasonable areas of the space, typically near the adjective itself or the corresponding noun (the noun *green* near *green N*), prototypical ANs for that adjective (*black face*), elements

related to the definition of the adjective (*mental activity*, *historical event*, *green colour*, *quick* and *little cost* for *easy N*), and so on.

American N	*black N*	*easy N*
Am. representative	black face	easy start
Am. territory	black hand	quick
Am. source	black (n)	little cost
green N	*historical N*	*mental N*
green (n)	historical	mental activity
red road	hist. event	mental experience
green colour	hist. content	mental energy
necessary N	*nice N*	*young N*
necessary	nice	youthful
necessary degree	good bit	young doctor
sufficient	nice break	young staff

Table 1: Nearest 3 neighbors of centroids of ANs that share the same adjective.

How about the neighbors of specific ANs? Table 2 reports the nearest 3 neighbors of 9 randomly selected ANs involving different adjectives (we inspected a larger random set, coming to similar conclusions to the ones emerging from this table).

bad luck	*electronic communication*	*historical map*
bad	elec. storage	topographical
bad weekend	elec. transmission	atlas
good spirit	purpose	hist. material
important route	*nice girl*	*little war*
important transport	good girl	great war
important road	big girl	major war
major road	guy	small war
red cover	*special collection*	*young husband*
black cover	general collection	small son
hardback	small collection	small daughter
red label	archives	mistress

Table 2: Nearest 3 neighbors of specific ANs.

The nearest neighbors of the corpus-based AN vectors in Table 2 make in general intuitive sense. Importantly, the neighbors pick up the composite meaning rather than that of the adjective or noun alone. For example, *cover* is an ambiguous word, but the *hardback* neighbor relates to its "front of a book" meaning that is the most natural one in combination with *red*. Similarly, it makes more sense that a *young husband* (rather than an old one) would have *small sons* and *daughters* (not to mention the *mistress*!).

We realize that the evidence presented here is of a very preliminary and intuitive nature. Indeed, we will argue in the next section that there are cases in which the corpus-derived AN vector might not be a good approximation to our semantic intuitions about the AN, and a model-composed AN vector is a better semantic surrogate. One of the most important avenues for further work will be to come to a better characterization of the behaviour of corpus-observed ANs, where they work and where the don't. Still, the neighbors of average and AN-specific vectors of Tables 1 and 2 suggest that, for the bulk of ANs, such corpus-based co-occurrence vectors are semantically reasonable.

6 Study 2: Predicting AN vectors

Having tentatively established that the sort of vectors we can harvest for ANs by directly collecting their corpus co-occurrences are reasonable representations of their composite meaning, we move on to the core question of whether it is possible to reconstruct the vector for an unobserved AN from information about its components. We use nearness to the corpus-observed vectors of held-out ANs as a very direct way to evaluate the quality of model-generated ANs, since we just saw that the observed ANs look reasonable (but see the caveats at the end of this section). We leave it to further work to assess the quality of the generated ANs in an applied setting, for example adapting Mitchell and Lapata's paraphrasing task to ANs. Since the observed vectors look like plausible representations of composite meaning, we expect that the closer the model-generated vectors are to the observed ones, the better they should also perform in any task that requires access to the composite meaning, and thus that the results of the current evaluation should correlate with applied performance.

More in detail, we evaluate here the composition methods (and the *adjective* and *noun* baselines) by computing, for each of them, the cosine of the test set AN vectors they generate (the "predicted" ANs) with the 41K vectors representing our extended vocabulary in semantic space, and looking at the position of the corresponding observed ANs (that were not used for training, in the supervised approaches)

in the cosine-ranked lists. The lower the rank, the better the approximation. For efficiency reasons, we flatten out the ranks after the top 1,000 neighbors.

The results are summarized in Table 3 by the median and the other quartiles, calculated across all 26,440 ANs in the test set. These measures (unlike mean and variance) are not affected by the cut-off after 1K neighbors. To put the reported results into perspective, a model with a first quartile rank of 999 does very significantly better than chance (the binomial probability of $1/4$ or more of 26,440 trials being successful with $\pi = 0.024$ is virtually 0, where the latter quantity is the probability of an observed AN being at rank 999 or lower according to a geometric distribution with $\pi = 1/40999$).

method	*25%*	*median*	*75%*
alm	17	170	≥1K
add	27	257	≥1K
noun	72	448	≥1K
mult	279	≥1K	≥1K
slm	629	≥1K	≥1K
adj	≥1K	≥1K	≥1K

Table 3: Quartile ranks of observed ANs in cosine-ranked lists of predicted AN neighbors.

Our proposed method, *alm*, emerges as the best approach. The difference with the second best model, *add* (the only other model that does better than the non-trivial baseline of using the component noun vector as a surrogate for AN), is highly statistically significant (Wilcoxon signed rank test, $p < 0.00001$). If we randomly downsample the AN set to keep an equal number of ANs per adjective (200), the difference is still significant with p below the same threshold, indicating that the general result is not due to a better performance of *alm* on a few common adjectives.[1]

Among the alternative models, the fact that the performance of *add* is decidedly better than that of *mult* is remarkable, since earlier studies found that multiplicative models are, in general, better than additive ones in compositionality tasks (see Section 2 above). This might depend on the nature of AN composition, but there are also more technical issues at hand: (i) we are not sure that previous studies normalized before summing like we did, and (ii) the multiplicative model, as discussed in Section 4, does not benefit from SVD reduction. The single linear mapping model (*slm*) proposed by Guevara (2010) is doing even worse than the multiplicative method, suggesting that a single set of weights does not provide enough flexibility to model a variety of adjective transformations successfully. This is at odds with Guevara's experiment in which *slm* outperformed *mult* and *add* on the task of ranking predicted ANs with respect to a target observed AN. Besides various differences in task definition and model implementation, Guevara trained his model on ANs that include a wide variety of adjectives, whereas our training data were limited to ANs containing one of our 36 test set adjectives. Future work should re-evalute the performance of Guevara's approach in our task, but under his training regime.

Looking now at the *alm* results in more detail, the best median ranks are obtained for very frequent adjectives. The top ones are *new* (median rank: 34), *great* (79), *American* (82), *large* (82) and *different* (97). There is a high inverse correlation between median rank and adjective frequency (Spearman's $\rho = -0.56$). Although from a statistical perspective it is expected that we get better results where we have more data, from a linguistic point of view it is interesting that *alm* works best with extremely frequent, highly polysemous adjectives like *new*, *large* and *different*, that border on function words – a domain where distributional semantics has generally not been tested.

Although, in relative terms and considering the difficulty of the task, *alm* performs well, it is still far from perfect – for 27% *alm*-predicted ANs, the observed vector is not even in the top 1K neighbor set! A qualitative look at some of the most problematic examples indicates however that a good proportion of them might actually not be instances where our model got the AN vector wrong, but cases of anomalous observed ANs. The left side of Table 4 compares the nearest neighbors (excluding each other) of the observed and *alm*-predicted vectors in 10 ran-

[1] The semantic space in which we rank the observed ANs with respect to their predicted counterparts also contain the observed vectors of nouns and ANs that were used to train *alm*. We do not see how this should affect performance, but we nevertheless repeated the evaluation leaving out, for each AN, the observed items used in training, and we obtained the same results reported in the main text (same ordering of method performance, and very significant difference between *alm* and *add*).

SIMILAR			DISSIMILAR		
adj N	*obs. neighbor*	*pred. neighbor*	*adj N*	*obs. neighbor*	*pred. neighbor*
common understanding	common approach	common vision	American affair	Am. development	Am. policy
different authority	diff. objective	diff. description	current dimension	left (a)	current element
different partner	diff. organisation	diff. department	good complaint	current complaint	good beginning
general question	general issue	*same*	great field	excellent field	gr. distribution
historical introduction	hist. background	*same*	historical thing	different today	hist. reality
necessary qualification	nec. experience	*same*	important summer	summer	big holiday
new actor	new cast	*same*	large pass	historical region	large dimension
recent request	recent enquiry	*same*	special something	little animal	special thing
small drop	droplet	drop	white profile	chrome (n)	white show
young engineer	young designer	y. engineering	young photo	important song	young image

Table 4: Left: nearest neighbors of observed and *alm*-predicted ANs (excluding each other) for a random set of ANs where rank of observed w.r.t. predicted is 1. Right: nearest neighbors of predicted and observed ANs for random set where rank of observed w.r.t. predicted is $\geq 1K$.

domly selected cases where the observed AN is the nearest neighbor of the predicted one. Here, the ANs themselves make sense, and the (often shared) neighbors are also sensible (*recent enquiry* for *recent request*, *common approach* and *common vision* for *common understanding*, etc.). Moving to the right, we see 10 random examples of ANs where the observed AN was at least 999 neighbors apart from the *alm* prediction. First, we notice some ANs that are difficult to interpret out-of-context (*important summer*, *white profile*, *young photo*, *large pass*, ...). Second, at least subjectively, we find that in many cases the nearest neighbor of predicted AN is actually more sensible than that of observed AN: *current element* (vs. *left*) for *current dimension*, *historical reality* (vs. *different today*) for *historical thing*, *special thing* (vs. *little animal*) for *special something*, *young image* (vs. *important song*) for *young photo*. In the other cases, the predicted AN neighbor is at least not obviously worse than the observed AN neighbor.

There is a high inverse correlation between the frequency of occurrence of an AN and the rank of the observed AN with respect to the predicted one ($\rho = -0.48$), suggesting that our model is worse at approximating the observed vectors of rare forms, that might, in turn, be those for which the corpus-based representation is less reliable. In these cases, dissimilarities between observed and expected vectors, rather than signaling problems with the model, might indicate that the predicted vector, based on a composition function learned from many examples, is *better* than the one directly extracted from the corpus. The examples in the right panel of Table 4 bring some preliminary support to this hypothesis, to be systematically explored in future work.

7 Study 3: Comparing adjectives

If adjectives are functions, and not corpus-derived vectors, is it still possible to compare them meaningfully? We explore two ways to accomplish this in our framework: one is to represent adjectives by the average of the AN vectors that contain them (the centroid vectors whose neighbors are illustrated in Table 1 above), and the other to compare them based on the 300×300 weight matrices we estimate from noun-AN pairs (we unfold these matrices into 90K-dimensional vectors). We compare the quality of these representations to that of the standard approach in distributional semantics, i.e., representing the adjectives directly with their corpus co-occurrence profile vectors (in our case, projected onto the SVD-reduced space).

We evaluate performance on the task of clustering those 19 adjectives in our set that can be relatively straightforwardly categorized into general classes comprising a minimum of 4 items. The test set built according to these criteria contains 4 classes: color (*white*, *black*, *red*, *green*), positive evaluation (*nice*, *excellent*, *important*, *major*, *appropriate*), time (*recent*, *new*, *current*, *old*, *young*), and size (*big*, *huge*, *little*, *small*, *large*). We cluster with the CLUTO toolkit (Karypis, 2003), using the *repeated bisections with global optimization*

method, accepting all of CLUTO's default values for this choice. Cluster quality is evaluated by percentage *purity* (Zhao and Karypis, 2003). If n_r^i is the number of items from the i-th true (gold standard) class assigned to the r-th cluster, n is the total number of items and k the number of clusters, then: $Purity = \frac{1}{n}\sum_{r=1}^{k}\max_i(n_r^i)$. We calculate empirical 95% confidence intervals around purity by a heuristic bootstrap procedure based on 10K resamplings of the data set (Efron and Tibshirani, 1994). The random baseline distribution is obtained by 10K random assignments of adjectives to the clusters, under the constraint that no cluster is empty.

Table 5 shows that all methods are significantly better than chance. Our two "indirect" representations achieve similar performance, and they are (slightly) better than the traditional method based on adjective co-occurrence vectors. We conclude that, although our approach does not provide a direct encoding of adjective meaning in terms of such independently collected vectors, it does have meaningful ways to represent their semantic properties.

input	*purity*
matrix	73.7 (68.4-94.7)
centroid	73.7 (63.2-94.7)
vector	68.4 (63.2-89.5)
random	45.9 (36.8-57.9)

Table 5: Percentage purity in adjective clustering with bootstrapped 95% confidence intervals.

8 Conclusion

The work we reported constitutes an encouraging start for our approach to modeling (AN) composition. We suggested, along the way, various directions for further studies. We consider the following issues to be the most pressing ones.

We currently train each adjective-specific model separately: We should explore hierarchical modeling approaches that exploit similarities across adjectives (and possibly syntactic constructions) to estimate better models.

Evaluation-wise, the differences between observed and predicted ANs must be analyzed more extensively, to support the claim that, when their vectors differ, model-based prediction improves on the observed vector. Evaluation in a more applied task should also be pursued – in particular, we will design a paraphrasing task similar to the one proposed by Mitchell and Lapata to evaluate noun-verb constructions.

Since we do not collect vectors for the "functor" component of a composition process (for AN constructions, the adjective), our approach naturally extends to processes that involve bound morphemes, such as affixation, where we would not need to collect independent co-occurrence information for the affixes. For example, to account for *re-* prefixation we do not need to collect a *re-* vector (required by all other approaches to composition), but simply vectors for a set of *V*/*reV* pairs, where both members of the pairs are words (e.g., *consider*/*reconsider*).

Our approach can also deal, out-of-the-box, with recursive constructions (*sad little red hat*), and can be easily extended to more abstract constructions, such as *determiner N* (mapping *dog* to *the/a/one dog*). Still, we need to design a good testing scenario to evaluate the quality of such model-generated constructions.

Ultimately, we want to compose larger and larger constituents, up to full sentences. It remains to be seen if the approach we proposed will scale up to such challenges.

Acknowledgments

We thank Gemma Boleda, Emilano Guevara, Alessandro Lenci, Louise McNally and the anonymous reviewers for useful information, advice and comments.

References

S. Clark and S. Pulman. 2007. Combining symbolic and distributional models of meaning. In *Proceedings of the First Symposium on Quantum Interaction*, pages 52–55.

B. Efron and R. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman and Hall, Boca Raton, FL.

K. Erk and S. Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP*, pages 897–906.

K. Erk and S. Padó. 2009. Paraphrase assessment in structured vector space: Exploring parameters and datasets. In *Proceedings of the EACL GEMS Workshop*, pages 57–65.

S. Evert. 2005. *The Statistics of Word Cooccurrences*. Dissertation, Stuttgart University.

P. Foltz, W. Kintsch, and Th. Landauer. 1998. The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25:285–307.

G. Frege. 1892. Über sinn und bedeutung. *Zeitschrift fuer Philosophie un philosophische Kritik*, 100.

E. Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the ACL GEMS Workshop*, pages 33–37.

T. Hastie, R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning, 2nd ed.* Springer, New York.

M. Jones and D. Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114:1–37.

H. Kamp. 1975. Two theories about adjectives. In E. Keenan, editor, *Formal Semantics of Natural Language*, pages 123–155. Cambridge University Press.

G. Karypis. 2003. CLUTO: A clustering toolkit. Technical Report 02-017, University of Minnesota Department of Computer Science.

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

Th. Landauer and S. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

B. Mevik and R. Wehrens. 2007. The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software*, 18(2).

J. Mitchell and M. Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL*, pages 236–244.

J. Mitchell and M. Lapata. 2009. Language models based on semantic composition. In *Proceedings of EMNLP*, pages 430–439.

R. Montague. 1970a. English as a formal language. In B. Visentini, editor, *Linguaggi nella Società e nella Tecnica*, pages 189–224. Edizioni di Comunità, Milan. Reprinted in Thomason (1974).

R. Montague. 1970b. Universal grammar. *Theoria*, 36:373–398. Reprinted in Thomason (1974).

R. Montague. 1973. The proper treatment of quantification in English. In K.J.J. Hintikka, editor, *Approaches to Natural Language*, pages 221–242. Reidel, Dordrecht. Reprinted in Thomason (1974).

B. Partee. 2004. Compositionality. In *Compositionality in Formal Semantics: Selected Papers by Barbara H. Partee*. Blackwell, Oxford.

R. Rapp. 2003. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the MT Summit*, pages 315–322.

H. Rubenstein and J. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

S. Rudolph and E. Giesbrecht. 2010. Compositional matrix-space models of language. In *Proceedings of ACL*.

M. Sahlgren, A. Holst, and P. Kanerva. 2008. Permutations as a means to encode order in word space. In *Proceedings of CogSci*, pages 1300–1305.

M. Sahlgren. 2006. *The Word-Space Model*. Dissertation, Stockholm University.

H. Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL-SIGDAT Workshop*.

H. Schütze. 1997. *Ambiguity Resolution in Natural Language Learning*. CSLI, Stanford, CA.

M. Siegel. 1976. *Capturing the Adjective*. Ph.D. thesis, University of Massachusetts at Amherst.

P. Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist networks. *Artificial Intelligence*, 46:159–216.

R. H. Thomason, editor. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New York.

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

D. Widdows. 2008. Semantic vector products: Some initial investigations. In *Proceedings of the Second Symposium on Quantum Interaction*, Oxford.

Y. Zhao and G. Karypis. 2003. Criterion functions for document clustering: Experiments and analysis. Technical Report 01-40, University of Minnesota Department of Computer Science.