(Linear) Maps of the Impossible: Capturing semantic anomalies in distributional space

Eva Maria Vecchi and **Marco Baroni** and **Roberto Zamparelli**
Center for Mind/Brain Sciences, University of Trento
Rovereto (TN), Italy
{evamaria.vecchi-1,marco.baroni,roberto.zamparelli}@unitn.it

Abstract

In this paper, we present a first attempt to characterize the semantic deviance of composite expressions in distributional semantics. Specifically, we look for properties of adjective-noun combinations within a vector-based semantic space that might cue their lack of meaning. We evaluate four different compositionality models shown to have various levels of success in representing the meaning of AN pairs: the simple additive and multiplicative models of Mitchell and Lapata (2008), and the linear-map-based models of Guevara (2010) and Baroni and Zamparelli (2010). For each model, we generate composite vectors for a set of AN combinations unattested in the source corpus and which have been deemed either *acceptable* or *semantically deviant*. We then compute measures that might cue semantic anomaly, and compare each model's results for the two classes of ANs. Our study shows that simple, unsupervised cues can indeed significantly tell unattested but acceptable ANs apart from impossible, or deviant, ANs, and that the simple additive and multiplicative models are the most effective in this task.

1 Introduction

Statistical approaches to describe, represent and understand natural language have been criticized as failing to account for linguistic 'creativity', a property which has been accredited to the compositional nature of natural language. Specifically, criticisms against statistical methods were based on the argument that a corpus cannot significantly sample a natural language because natural language is infinite (Chomsky, 1957). This cricticism also applies to distributional semantic models that build semantic representations of words or phrases in terms of vectors recording their distributional co-occurrence patterns in a corpus (Turney and Pantel, 2010), but have no obvious way to generalize to word combinations that have not been observed in the corpus. To address this problem, there have been several recent attempts to incorporate into distributional semantic models a component that generates vectors for unseen linguistic structures by compositional operations in the vector space (Baroni and Zamparelli, 2010; Guevara, 2010; Mitchell and Lapata, 2010).

The ability to work with unattested data leads to the question of why a linguistic expression might not be attested in even an extremely large and well-balanced corpus. Its absence might be motivated by a number of factors: pure chance, the fact that the expression is ungrammatical, uses a rare structure, describes false facts, or, finally, is *nonsensical*. One criticism from generative linguists is precisely that statistical methods could not distinguish between these various possibilities.

The difficulty of solving this problem can be illustrated by the difference in semantics between the adjective-noun pairs in (1a) and (1b):

(1) a. blue rose
b. residential steak

Although it may be the case that you have never ac-

Proceedings of the Workshop on Distributional Semantics and Compositionality (DiSCo'2011), pages 1–9,

tually seen a *blue rose*, the concept is not inconceivable. On the other hand, the concept of a *residential steak* is rather unimaginable, and intuitively its absence in a corpus is motivated by more than just chance or data sparseness.

The present paper is a first attempt to use compositionality and distributional measures to distinguish nonsensical, or semantically deviant, linguistic expression from other types of unattested structures. The task of distinguishing between unattested but **acceptable** and unattested but **semantically deviant** linguistic expressions is not only a way to address the criticism about the meaning of 'unattestedness', but also a task that could have a large impact on the (computational) linguistic community as a whole (see Section 2.1).

Our specific goal is to automatically detect semantic deviance in attributive Adjective-Noun (AN) expressions, using a small number of simple cues in the vectorial representation of an AN as it is generated from the distributional vectors of its component A and N by four compositional models found in the literature. The choice of AN as our testbed is motivated by two facts: first of all, ANs are common, small constituents containing no functional material, and secondly, ANs have already been studied in compositional distributional semantics (Baroni and Zamparelli, 2010; Guevara, 2010; Mitchell and Lapata, 2010).

It is important to note that in this research we talk about 'semantically deviant' expressions, but we do not exclude the possibility that such expressions are interpreted as metaphors, via a chain of associations. In fact, distributional measures are desirable models to account for this, since they naturally lead to a gradient notion of semantic anomaly.

The rest of this paper is structured as follows. Section 2 discusses relevant earlier work, introducing the literature on semantic deviance as well as compositional methods in distributional semantics. Section 3 presents some hypotheses about cues of semantic deviance in distributional space. Our experimental setup and procedure are detailed in Section 4, whereas the experiments' results are presented and analyzed in Section 5. We conclude by summarizing and proposing future directions in Section 6.

2 Related work

2.1 Semantic deviance

As far as we know, we are the first to try to model semantic deviance using distributional methods, but the issue of when a complex linguistic expression is semantically deviant has been addressed since the 1950's in various areas of linguistics. In computational linguistics, the possibility of detecting semantic deviance has been seen as a prerequisite to access metaphorical/non-literal semantic interpretations (Fass and Wilks, 1983; Zhou et al., 2007). In psycholinguistics, it has been part of a wide debate on the point at which context can make us perceive a 'literal' vs. a 'figurative' meaning (Giora, 2002). In theoretical generative linguistics, the issue was originally part of a discussion on the boundaries between syntax and semantics. Cases like Chomsky's classic "*Colorless green ideas sleep furiously*" can actually be regarded as violations of very fine-grained syntactic *selectional restrictions* on the arguments of verbs or modifiers, on the model of **much computer* (arguably a failure of *much* to combine with a noun +COUNT). By 1977, even Chomsky doubted that speakers could in general have intuitions about whether ill-formedness was syntactic or semantic (Chomsky, 1977, p. 4). The spirit of the selectional approach persists in Asher (2011), who proposes a detailed system of semantic types plus a theory of type coercion, designed to account for the shift in meaning seen in, e.g., (2) (*lunch* as food or as an event).

(2) Lunch was delicious but took forever.

A practical problem with this approach is that a full handmade specification of the features that determine semantic compatibility is a very expensive and time-consuming enterprise, and it should be done consistently across the whole content lexicon. Moreover, it is unclear how to model the intuition that *naval fraction*, *musical North* or *institutional acid* sound odd, in the absence of very particular contexts, while (2) sounds quite natural. Whatever the nature of coercion, we do not want it to run so smoothly that any combination of A and N (or V and its arguments) becomes meaningful and completely acceptable.

2.2 Distributional approaches to meaning composition

Although the issue of how to compose meaning has attracted interest since the early days of distributional semantics (Landauer and Dumais, 1997), recently a very general framework for modeling compositionality has been proposed by Mitchell and Lapata (Mitchell and Lapata, 2008; Mitchell and Lapata, 2009; Mitchell and Lapata, 2010). Given two vectors $\mathbf{u}$ and $\mathbf{v}$, they identify two general classes of composition models, (linear) additive models:

$$\mathbf{p} = \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} \quad (1)$$

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

$$\mathbf{p} = \mathbf{C}\mathbf{u}\mathbf{v}$$

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: The simplified additive model given by $\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v}$, and a simplified multiplicative approach that reduces to component-wise multiplication, where the i-th component of the composed vector is given by: $p_i = u_i v_i$. Mitchell and Lapata evaluate the simplified models on a wide range of tasks ranging from paraphrasing to statistical language modeling to predicting similarity intuitions. Both simple models fare quite well across tasks and alternative semantic representations, also when compared to more complex methods derived from the equations above. Given their overall simplicity, good performance and the fact that they have also been extensively tested in other studies (Baroni and Zamparelli, 2010; Erk and Padó, 2008; Guevara, 2010; Kintsch, 2001; Landauer and Dumais, 1997), we re-implement here both the simplified additive and simplified multiplicative methods (we do not, however, attempt to tune the weights of the additive model, although we do apply a scalar normalization constant to the adjective and noun vectors).

Mitchell and Lapata (as well as earlier 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 corpus-observed ANs to train a supervised composition model. Guevara, whose approach we also re-implement here, adopts the full additive composition form from Equation (1) and he estimates the $\mathbf{A}$ and $\mathbf{B}$ weights (concatenated into a single matrix, that acts as a linear map from the space of concatenated adjective and noun vectors onto the AN vector space) 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. Corpus-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.

The final approach we re-implement is the one proposed by Baroni and Zamparelli (2010), who treat attributive adjectives as functions from noun meanings to noun meanings. This is a standard approach in Montague semantics (Thomason, 1974), except noun meanings here are distributional vectors, not denotations, and adjectives are (linear) functions learned from a large corpus. Unlike in Guevara's approach, a separate matrix is generated for each adjective using only examples of ANs containing that adjective, and no adjective vector is used: the adjective is represented entirely by the matrix mapping nouns to ANs. In terms of Mitchell and Lapata's general framework, this approach derives from the additive form in Equation (1) with the matrix multiplying the adjective vector (say, $\mathbf{A}$) set to $\mathbf{0}$, the other matrix ($\mathbf{B}$) representing the adjective at hand, and $\mathbf{v}$ a noun vector. Baroni and Zamparelli (2010) show that their model significantly outperforms other vector composition methods, including addition, multiplication and Guevara's approach, in the task of approximating the correct vectors for previously unseen (but corpus-attested) ANs. Simple addition emerges as the second best model.

See Section 4.3 below for details on our re-implementations. Note that they follow very closely the procedure of Baroni and Zamparelli (2010), including choices of source corpus and parameter values, so that we expect their results on the quality of the various models in predicting ANs to also hold for our re-implementations.

3 Simple indices of semantic deviance

We consider here a few simple, unsupervised measures to help us distinguish the representation that a distributional composition model generates for a semantically anomalous AN from the one it generates for a semantically acceptable AN. In both cases, we assume that the AN is not already part of the model semantic space, just like you can distinguish between *parliamentary tomato* (odd) and *marble iPad* (OK), although you probably never heard either expression.

We hypothesize that, since the values in the dimensions of a semantic space are a distributional proxy to the meaning of an expression, a meaningless expression should in general have low values across the semantic space dimensions. For example, a *parliamentary tomato*, no longer being a vegetable but being an unlikely parliamentary event, might have low values on both dimensions characterizing vegetables and dimensions characterizing events. Thus, our first simple measure of semantic anomaly is the **length** of the model-generated AN. We hypothesize that anomalous AN vectors are shorter than acceptable ANs.

Second, if deviant composition destroys or randomizes the meaning of a noun, as a side effect we might expect the resulting AN to be more distant, in the semantic space, from the component noun. Although even a *marble iPad* might have lost some essential properties of iPads (it could for example be an iPad statue you cannot use as a tablet), to the extent that we can make sense of it, it must retain at least some characteristics of iPads (at the very least, it will be shaped like an iPad). On the other hand, we cannot imagine what a *parliamentary tomato* should be, and thus cannot attribute even a subset of the regular tomato properties to it. We thus hypothesize that model-generated vectors of deviant ANs will form a wider angle (equivalently, will have a lower **cosine**) with the corresponding N vectors than acceptable ANs.

Finally, if an AN makes no sense, its model-generated vector should not have many neighbours in the semantic space, since our semantic space is populated by nouns, adjectives and ANs that are commonly encountered in the corpus, and should thus be meaningful. We expect deviant ANs to be "semantically isolated", a notion that we operationalize in terms of a (neighborhood) **density** measure, namely the average cosine with the (top 10) nearest neighbours. We hypothesize that model-generated vectors of deviant ANs will have lower density than model-generated acceptable ANs.

4 Experimental setup

4.1 Semantic space

Our initial step was to construct a semantic space for our experiments, consisting of a matrix where each row vector represents an adjective, noun or AN. We first introduce the source corpus, then the vocabulary of words and ANs that we represent in the space, and finally the procedure adopted to build the vectors representing the vocabulary items from corpus statistics, in order to obtain the semantic space matrix. We work here with a "vanilla" semantic space (essentially, we follow the steps of Baroni and Zamparelli (2010)), since our focus is on the effect of different composition methods given a common semantic space. We leave it to further work to study how choices in semantic space construction affect composition operations.

4.1.1 Source corpus

We use as our source corpus the concatenation of 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/`). The corpus has been tokenized, POS-tagged and lemmatized with the TreeTagger (Schmid, 1995), and it contains about 2.8 billion tokens. We extract all statistics at the lemma level, ignoring inflectional information.

4.1.2 Semantic space vocabulary

The words/ANs in the semantic space must of course include the items that we need for our experiments (adjectives, nouns and ANs used for model training and as input to composition). Moreover, in order to study the behaviour of the test items we are interested in (that is, model-generated AN vectors) within a large and less ad-hoc space, we also include many more adjectives, nouns and ANs in our vocabulary not directly relevant to our experimental manipulations.

We populate our semantic space with the 8K most frequent nouns and 4K most frequent adjectives from the corpus (excluding, in both cases, the top 50 most frequent elements). We extended this vocabulary to include two sets of ANs (33K ANs cumulatively), for a total of 45K vocabulary items in the semantic space.

To create the ANs needed to run and evaluate the experiments described below, we focused on a set of adjectives which are very frequent in the corpus so that they will be in general able to combine with wide classes of nouns, making the unattested cases more interesting, but not so frequent as to have such a general meaning that would permit a free combination with nearly any noun. The ANs were therefore generated by crossing a selected set of 200 very frequent adjectives (adjectives attested in the corpus at least 47K times, and at most 740K) and the set of the 8K nouns in our semantic space vocabulary, producing a set of 4.92M generated ANs.

The first set of ANs included in the semantic space vocabulary is a randomly sampled set of 30K ANs from the generated set which are attested in the corpus at least 200 times (to avoid noise and focus on ANs for which we can extract reasonably robust distributional data). We also extracted any unattested ANs from the set of generated set (about 3.5M unattested ANs), putting them aside to later assemble our evaluation material, described in Section 4.2.

To add further variety to the semantic space, we included a less controlled second set of 3K ANs randomly picked among those that are attested and are formed by the combination of any of the 4K adjectives and 8K nouns in the vocabulary.

4.1.3 Semantic space construction

For each of the items in our vocabulary, we first build 10K-dimensional vectors by recording their sentence-internal co-occurrence with the top 10K most frequent content words (nouns, adjectives or verbs) in the corpus. The raw co-occurrence counts are then transformed into Local Mutual Information scores (Local Mutual Information is an association measure that closely approximates the commonly used Log-Likelihood Ratio while being simpler to compute (Baroni and Lenci, 2010; Evert, 2005)).

Next, we reduce the full co-occurrence matrix applying the Singular Value Decomposition (SVD) operation, like in LSA and related distributional semantic methods (Landauer and Dumais, 1997; Rapp, 2003; Schütze, 1997). The original 45K-by-10K-dimensional matrix is reduced in this way to a 45K-by-300 matrix, where vocabulary items are represented by their coordinates in the space spanned by the first 300 right singular vectors of the SVD solution. This step is motivated by the fact that we will estimate linear models to predict the values of each dimension of an AN from the dimensions of the components. We thus prefer to work in a smaller and denser space. As a sanity check, we verify that we obtain state-of-the-art-range results on various semantic tasks using this reduced semantic space (not reported here for space reason).

4.2 Evaluation materials

Our goal is to study what happens when compositional methods are used to construct a distributional representation for ANs that are semantically deviant, compared to the AN representations they generate for ANs they have not encountered before, but that are semantically acceptable.

In order to assemble these lists, we started from the set of 3.5M unattested ANs described in Section 4.1.2 above, focusing on 30 randomly chosen adjectives. For each of these, we randomly picked 100 ANs for manual inspection (3K ANs in total). Two authors went through this list, marking those ANs that they found semantically highly anomalous, no matter how much effort one would put in constructing metaphorical or context-dependent interpretations, as well as those they found completely acceptable (so, rating was on a 3-way scale: deviant,

intermediate, acceptable). The rating exercise resulted in rather low agreement (Cohen's $\kappa = 0.32$), but we reasoned that those relatively few cases (456 over 3K) where both judges agreed the AN was odd should indeed be odd, and similarly for the even rarer cases in which they agreed an AN was completely acceptable (334 over 3K). We thus used the agreed deviant and acceptable ANs as test data.

Of 30 adjectives, 5 were discarded for either technical reasons or for having less than 5 agreed deviant or acceptable ANs. This left us with a **deviant AN test set** comprising of 413 ANs, on average 16 for each of the 25 remaining adjectives. Some examples of ANs in this set are: *academic bladder*, *blind pronunciation*, *parliamentary potato* and *sharp glue*. The **acceptable** (but unattested) **AN test set** contains 280 ANs, on average 11 for each of the 25 studied adjectives. Examples of ANs in this set include: *vulnerable gunman*, *huge joystick*, *academic crusade* and *blind cook*. The evaluation sets can be downloaded from `http://www.vecchi.com/eva/resources.html`.

There is no significant difference between the length of the vectors of the component nouns in the acceptable vs. deviant AN sets (two-tailed Welch's t test; $t = -0.25$; $p > 0.8$). This is important, since at least one of the potential cues to deviance we consider (AN vector length) is length-dependent, and we do not want a trivial result that can simply be explained by systematic differences in the length of the input vectors.

4.3 Composition methods

As discussed in Section 2.2, the experiment was carried out across four compositional methods.

Additive AN vectors (*add* method) are simply obtained by summing the corresponding adjective and noun vectors after normalizing them. **Multiplicative** vectors (*mult* method) were obtained by component-wise multiplication of the adjective and noun vectors, also after normalization. Confirming the results of Baroni and Zamparelli (2010), non-normalized versions of *add* and *mult* were also tested, but did not produce significant results (in the case of multiplication, normalization amounts to multiplying the composite vector by a scalar, so it only affects the length-dependent vector length measure). It is important to note that, as reported in Baroni and Zamparelli (2010), the *mult* method can be expected to perform better in the original, non-reduced semantic space 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 instead of being cancelled out). The tests of Section 5, however, are each run in the SVD-reduced space to remain consistent across all models. We leave it to future work to explore the effect on the performance of using the non-reduced space for the models for which this option is computationally viable.

In the **linear map** (*lm*) approach proposed by Guevara (2010), a composite AN vector is obtained by multiplying a weight matrix by the concatenation of the adjective and noun vectors, so that each dimension of the generated AN vector is a linear combination of dimensions of the corresponding adjective and noun vectors. That is, the 600 weights in each of the 300 rows of the weight matrix are the coefficients of a linear equation predicting the values of a single dimension in the AN vector as a linear combination (weighted sum) of the 300 adjective and 300 noun dimensions. Following Guevara, we estimate the coefficients of the equation using (multivariate) partial least squares regression (PLSR) as implemented in the R `pls` package (Mevik and Wehrens, 2007), with the latent dimension parameter of PLSR set to 50, the same value used by Baroni and Zamparelli (2010). Coefficient matrix estimation is performed by feeding the PLSR a set of input-output examples, where the input is given by concatenated adjective and noun vectors, and the output is the vector of the corresponding AN directly extracted from our semantic space (i.e., the AN vectors used in training are not model-generated, but directly derived from corpus evidence about their distribution). The matrix is estimated using a random sample of 2K adjective-noun-AN tuples where the AN belongs to the set of 30K frequently attested ANs in our vocabulary.

Finally, in the **adjective-specific linear map** (*alm*) method of Baroni and Zamparelli (2010), an AN is generated by multiplying an adjective weight matrix with a noun vector. The weights of each of the 300 rows of the weight matrix are the coefficients of a linear equation predicting the values of one of

the dimensions of the AN vector as a linear combination of the 300 dimensions of the component noun. The linear equation coefficients are estimated separately for each of the 25 tested adjectives from the attested noun-AN pairs containing that adjective (observed adjective vectors are not used), again using PLSR with the same parameter as above. For each adjective, the training N-AN vector pairs chosen are those available in the semantic space for each test set adjective, and range from 100 to more than 500 items across the 25 adjectives.

4.4 Experimental procedure

Using each composition method, we generate composite vectors for all the ANs in the two (acceptable and deviant) evaluation sets (see Section 4.2 above). We then compute the measures that might cue semantic deviance discussed in Section 3 above, and compare their values between the two AN sets. In order to smooth out adjective-specific effects, we z-normalize the values of each measure across all the ANs sharing an adjective before computing global statistics (i.e., the values for all ANs sharing an adjective from the two sets are transformed by subtracting their mean and dividing by their variance). We then compare the two sets, for each composition method and deviance cue, by means of two-tailed Welch's t tests. We report the estimated t score, that is, the standardized difference between the mean acceptable and deviant AN values, with the corresponding significance level. For all our cues, we predict t to be significantly larger than 0: Acceptable AN vectors should be *longer* than deviant ones, they should be *nearer* – that is, have a higher cosine with – the component N vectors and their neighbourhood should be *denser* – that is, the average cosines with their top neighbours should be higher than the ones of deviant ANs with their top neighbours.

5 Results

The results of our experiments are summarized in Table 1. We see that *add* and *mult* provide significant results in the expected direction for 2 over 3 cues, only failing the cosine test. With the *lm* model, acceptable and deviant ANs are indistinguishable across the board, whereas *alm* captures the distinction in terms of density.

	LENGTH		COSINE		DENSITY	
method	t	*sig.*	t	*sig.*	t	*sig.*
add	7.89	*	0.31		2.63	*
mult	3.16	*	-0.56		2.68	*
lm	0.16		0.55		-0.23	
alm	0.48		1.37		3.12	*

Table 1: t scores for difference between acceptable and deviant ANs with respect to 3 cues of deviance: *length* of the AN vector, *cosine* of the AN vector with the component noun vector and *density*, measured as the average cosine of an AN vector with its nearest 10 neighbours in semantic space. For all significant results, $p< 0.01$.

The high scores in the vector length analyses of both the addition and the multiplication models are an indication that semantically acceptable ANs tend to be composed of *similar* adjectives and nouns, i.e., those which occur in similar contexts and we can assume are likely to belong to the same domain, which sounds plausible.

In Baroni and Zamparelli (2010), the *alm* model performed far better than *add* and *mult* in approximating the correct vectors for unseen ANs, while on this (in a sense, more metalinguistic) task *add* and *mult* work better, while *alm* is successful only in the more sophisticated measure of neighbor density.

The lack of significant results for the cosine measure is disappointing, but not entirely surprising. A large angle between N and AN might be a feature of impossible ANs common to various types of possible ANs: idioms (a *red herring* is probably far from *herring* in semantic space), non-subsective adjectives (*stone lion* vs. *lion*; *fake butterfly* vs. *butterfly*), plus some metaphorical constructions (*academic crusade* vs. *crusade*—one of several ANs judged acceptable in our study, which can only be taken as metaphors). Recall, finally, that the vector for the base N collapses together all the meanings of an ambiguous N. The adjective might have a disambiguating effect which would increase the cosine distance.

To gain a better understanding of the neighborhood density test we performed a detailed analysis of the nearest neighbors of the AN vectors generated by the three models in which the difference in neighbor distance was significant across deviant and acceptable ANs: *alm*, multiplication and addition. For

each of the ANs, we looked at the top 10 semantic-space neighbors generated by each of the three models, focusing on two aspects: whether the neighbor was a single A or N, rather than AN, and whether the neighbor contained the same A or N as the AN is was the neighbor of (as in *blind regatta / blind athlete* or *biological derivative / partial derivative*). The results are summarized in Table 2.

method	*status*	A *only*	N *only*	A_1= A_2	N_1= N_2
add	accept	11.9	8.7	14.6	2.4
	deviant	12.5	6.8	14.6	2.3
mult	accept	6.9	8.0	0.7	0.1
	deviant	2.7	7.3	0.5	0.1
alm	accept	4.9	17.7	7.0	0.0
	deviant	7.1	19.6	6.2	0.0

Table 2: Percentage distributions of various properties of the top 10 neighbours of ANs in the acceptable (2800) and deviant (4130) sets for *add*, *mult* and *alm*. The last two columns express whether the neighbor contains the same Adjective or Noun as the target AN.

In terms of the properties we measured, neighbor distributions are quite similar across acceptable and deviant ANs. One interesting finding is that the system is quite 'adjective-driven': particularly for the additive model (where we can imagine that some Ns with low dimensional values do not shift much the adjective position in the multidimensional space), less so in the *alm* method, and not at all for *mult*. To put the third and forth columns in context, the subset of the semantic space used to generate the SVD from which the neighbors are drawn contained 2.69% adjectives, 5.24% nouns and 92.07% ANs. With respect to the last two columns, it is interesting to observe that matching As are frequent for deviant ANs even in *alm*, a model which has never seen A-vectors during training. Further qualitative evaluations show that in many deviant AN cases the similarity is between the A in the target AN and the N of the neighbor (e.g. *academic bladder / honorary lectureship*), while the opposite effect seems to be much harder to find.

6 Conclusion and future work

The main aim of this paper was to propose a new challenge to the computational distributional semantics community, namely that of characterizing what happens, distributionally, when composition leads to semantically anomalous composite expressions. The hope is, on the one hand, to bring further support to the distributional approach by showing that it can be both productive and constrained; and on the other, to provide a more general characterization of the somewhat elusive notion of semantic deviance – a notion that the field of formal semantics acknowledges but might lack the right tools to model.

Our results are very preliminary, but also very encouraging, suggesting that simple unsupervised cues can significantly tell unattested but acceptable ANs apart from impossible, or at least deviant, ones. Although, somewhat disappointingly, the model that has been shown in a previous study (Baroni and Zamparelli, 2010) to be the best at capturing the semantics of well-formed ANs turns out to be worse than simple addition and multiplication.

Future avenues of research must include, first of all, an exploration on the effect on each model when tested in the non-reduced space where computationally possible, or using different dimensionality reduction methods. A preliminary study demonstrates an enhanced performance of the *mult* method in the full space.

Second, we hope to provide a larger benchmark of acceptable and deviant ANs, beyond the few hundreds we used here, and sampling a larger typology of ANs across frequency ranges and adjective and noun classes. To this extent, we are implementing a crowd-sourcing study to collect human judgments from a large pool of speakers on a much larger set of ANs unattested in the corpus. Averaging over multiple judgments, we will also be able to characterize semantic deviance as a gradient property, probably more accurately.

Next, the range of cues we used was quite limited, and we intend to extend the range to include more sophisticated methods such as 1) combining multiple cues in a single score; 2) training a supervised classifier from labeled acceptable and deviant ANs, and studying the most distinctive features discovered by the classifier; 3) trying more complex unsupervised techniques, such as using graph-theoretical methods to characterize the semantic neighborhood of ANs beyond our simple density measure.

Finally, we are currently not attempting a typol-

ogy of deviant ANs. We do not distinguish cases such as *parliamentary tomato*, where the adjective does not apply to the conceptual semantic type of the noun (or at least, where it is completely undetermined which relation could bridge the two objects), from oxymorons such as *dry water*, or vacuously redundant ANs (*liquid water*) and so on. We realize that, at a more advanced stage of the analysis, some of these categories might need to be explicitly distinguished (for example, *liquid water* is odd but perfectly meaningful), leading to a multi-way task. Similarly, among acceptable ANs, there are special classes of expressions, such as idiomatic constructions, metaphors or other rhetorical figures, that might be particularly difficult to distinguish from deviant ANs. Again, more cogent tasks involving such well-formed but non-literal constructions (beyond the examples that ended up by chance in our acceptable set) are left to future work.

Acknowledgments

We thank Raffaella Bernardi, Gemma Boleda, Louise McNally and the anonymous reviewers for their advice and comments.

References

Nicholas Asher. 2011. *Lexical Meaning in Context: A Web of Words*. Cambridge University Press.

Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, pages 1183–1193, Boston, MA.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton.

Noam Chomsky. 1977. *Essays on Form and Interpretation*. North Holland, New York.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP*, pages 897–906, Honolulu, HI, USA.

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

Dan Fass and Yorick Wilks. 1983. Preference semantics, ill-formedness, and metaphor. *Computational Linguistics*, 9:178–187.

Rachel Giora. 2002. Literal vs. figurative language: Different or equal? *Journal of Pragmatics*, 34:487–506.

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

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

Thomas Landauer and Susan 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.

Björn-Helge Mevik and Ron Wehrens. 2007. The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software*, 18(2). Published online: `http://www.jstatsoft.org/v18/i02/`.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL*, pages 236–244, Columbus, OH, USA.

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

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*.

Reinhard Rapp. 2003. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the 9th MT Summit*, pages 315–322, New Orleans, LA, USA.

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

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

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

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

Chang-Le Zhou, Yun Yang, and Xiao-Xi Huang. 2007. Computational mechanisms for metaphor in languages: a survey. *Journal of Computer Science and Technology*, 22:308–319.