Vector-based Models of Semantic Composition

Jeff Mitchell and **Mirella Lapata**
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, UK
jeff.mitchell@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

This paper proposes a framework for representing the meaning of phrases and sentences in vector space. Central to our approach is vector composition which we operationalize in terms of additive and multiplicative functions. Under this framework, we introduce a wide range of composition models which we evaluate empirically on a sentence similarity task. Experimental results demonstrate that the multiplicative models are superior to the additive alternatives when compared against human judgments.

1 Introduction

Vector-based models of word meaning (Lund and Burgess, 1996; Landauer and Dumais, 1997) have become increasingly popular in natural language processing (NLP) and cognitive science. The appeal of these models lies in their ability to represent meaning simply by using distributional information under the assumption that words occurring within similar contexts are semantically similar (Harris, 1968).

A variety of NLP tasks have made good use of vector-based models. Examples include automatic thesaurus extraction (Grefenstette, 1994), word sense discrimination (Schütze, 1998) and disambiguation (McCarthy et al., 2004), collocation extraction (Schone and Jurafsky, 2001), text segmentation (Choi et al., 2001) , and notably information retrieval (Salton et al., 1975). In cognitive science vector-based models have been successful in simulating semantic priming (Lund and Burgess, 1996; Landauer and Dumais, 1997) and text comprehension (Landauer and Dumais, 1997; Foltz et al., 1998). Moreover, the vector similarities within such semantic spaces have been shown to substantially correlate with human similarity judgments (McDonald, 2000) and word association norms (Denhire and Lemaire, 2004).

Despite their widespread use, vector-based models are typically directed at representing words in isolation and methods for constructing representations for phrases or sentences have received little attention in the literature. In fact, the commonest method for combining the vectors is to average them. Vector averaging is unfortunately insensitive to word order, and more generally syntactic structure, giving the same representation to any constructions that happen to share the same vocabulary. This is illustrated in the example below taken from Landauer et al. (1997). Sentences (1-a) and (1-b) contain exactly the same set of words but their meaning is entirely different.

(1) a. It was not the sales manager who hit the bottle that day, but the office worker with the serious drinking problem.
b. That day the office manager, who was drinking, hit the problem sales worker with a bottle, but it was not serious.

While vector addition has been effective in some applications such as essay grading (Landauer and Dumais, 1997) and coherence assessment (Foltz et al., 1998), there is ample empirical evidence that syntactic relations across and within sentences are crucial for sentence and discourse processing (Neville et al., 1991; West and Stanovich, 1986) and modulate cognitive behavior in sentence priming (Till et al., 1988) and inference tasks (Heit and

Proceedings of ACL-08: HLT, pages 236–244,
Columbus, Ohio, USA, June 2008.

Rubinstein, 1994).

Computational models of semantics which use symbolic logic representations (Montague, 1974) can account naturally for the meaning of phrases or sentences. Central in these models is the notion of compositionality — the meaning of complex expressions is determined by the meanings of their constituent expressions and the rules used to combine them. Here, semantic analysis is guided by syntactic structure, and therefore sentences (1-a) and (1-b) receive distinct representations. The downside of this approach is that differences in meaning are qualitative rather than quantitative, and degrees of similarity cannot be expressed easily.

In this paper we examine models of semantic composition that are empirically grounded and can represent similarity relations. We present a general framework for vector-based composition which allows us to consider different classes of models. Specifically, we present both additive and multiplicative models of vector combination and assess their performance on a sentence similarity rating experiment. Our results show that the multiplicative models are superior and correlate significantly with behavioral data.

2 Related Work

The problem of vector composition has received some attention in the connectionist literature, particularly in response to criticisms of the ability of connectionist representations to handle complex structures (Fodor and Pylyshyn, 1988). While neural networks can readily represent single distinct objects, in the case of multiple objects there are fundamental difficulties in keeping track of which features are bound to which objects. For the hierarchical structure of natural language this binding problem becomes particularly acute. For example, simplistic approaches to handling sentences such as *John loves Mary* and *Mary loves John* typically fail to make valid representations in one of two ways. Either there is a failure to distinguish between these two structures, because the network fails to keep track of the fact that *John* is subject in one and object in the other, or there is a failure to recognize that both structures involve the same participants, because *John* as a subject has a distinct representation from *John* as an object. In contrast, symbolic representations can naturally handle the binding of constituents to their roles, in a systematic manner that avoids both these problems.

Smolensky (1990) proposed the use of tensor products as a means of binding one vector to another. The tensor product $\mathbf{u} \otimes \mathbf{v}$ is a matrix whose components are all the possible products $u_i v_j$ of the components of vectors $\mathbf{u}$ and $\mathbf{v}$. A major difficulty with tensor products is their dimensionality which is higher than the dimensionality of the original vectors (precisely, the tensor product has dimensionality $m \times n$). To overcome this problem, other techniques have been proposed in which the binding of two vectors results in a vector which has the same dimensionality as its components. Holographic reduced representations (Plate, 1991) are one implementation of this idea where the tensor product is projected back onto the space of its components.

The projection is defined in terms of *circular convolution* a mathematical function that compresses the tensor product of two vectors. The compression is achieved by summing along the transdiagonal elements of the tensor product. Noisy versions of the original vectors can be recovered by means of *circular correlation* which is the approximate inverse of circular convolution. The success of circular correlation crucially depends on the components of the n-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$ being randomly distributed with mean 0 and variance $\frac{1}{n}$. This poses problems for modeling linguistic data which is typically represented by vectors with non-random structure.

Vector addition is by far the most common method for representing the meaning of linguistic sequences. For example, assuming that individual words are represented by vectors, we can compute the meaning of a sentence by taking their mean (Foltz et al., 1998; Landauer and Dumais, 1997). Vector addition does not increase the dimensionality of the resulting vector. However, since it is order independent, it cannot capture meaning differences that are modulated by differences in syntactic structure. Kintsch (2001) proposes a variation on the vector addition theme in an attempt to model how the meaning of a predicate (e.g., *run*) varies depending on the arguments it operates upon (e.g, *the horse ran* vs. *the color ran*). The idea is to add not only the vectors representing the predicate and its argument but also the neighbors associated with both of them. The neighbors, Kintsch argues, can ‘strengthen features of the predicate that are appropriate for the argument of the predication’.

	animal	stable	village	gallop	jokey
horse	0	6	2	10	4
run	1	8	4	4	0

Figure 1: A hypothetical semantic space for *horse* and *run*

Unfortunately, comparisons across vector composition models have been few and far between in the literature. The merits of different approaches are illustrated with a few hand picked examples and parameter values and large scale evaluations are uniformly absent (see Frank et al. (2007) for a criticism of Kintsch's (2001) evaluation standards). Our work proposes a framework for vector composition which allows the derivation of different types of models and licenses two fundamental composition operations, multiplication and addition (and their combination). Under this framework, we introduce novel composition models which we compare empirically against previous work using a rigorous evaluation methodology.

3 Composition Models

We formulate semantic composition as a function of two vectors, **u** and **v**. We assume that individual words are represented by vectors acquired from a corpus following any of the parametrisations that have been suggested in the literature.[1] We briefly note here that a word's vector typically represents its co-occurrence with neighboring words. The construction of the semantic space depends on the definition of linguistic context (e.g., neighbouring words can be documents or collocations), the number of components used (e.g., the k most frequent words in a corpus), and their values (e.g., as raw co-occurrence frequencies or ratios of probabilities). A hypothetical semantic space is illustrated in Figure 1. Here, the space has only five dimensions, and the matrix cells denote the co-occurrence of the target words (*horse* and *run*) with the context words *animal*, *stable*, and so on.

Let **p** denote the composition of two vectors **u** and **v**, representing a pair of constituents which stand in some syntactic relation R. Let K stand for any additional knowledge or information which is needed to construct the semantics of their composition. We define a general class of models for this process of composition as:

$$\mathbf{p} = f(\mathbf{u},\mathbf{v},R,K) \tag{1}$$

The expression above allows us to derive models for which **p** is constructed in a distinct space from **u** and **v**, as is the case for tensor products. It also allows us to derive models in which composition makes use of background knowledge K and models in which composition has a dependence, via the argument R, on syntax.

To derive specific models from this general framework requires the identification of appropriate constraints to narrow the space of functions being considered. One particularly useful constraint is to hold R fixed by focusing on a single well defined linguistic structure, for example the verb-subject relation. Another simplification concerns K which can be ignored so as to explore what can be achieved in the absence of additional knowledge. This reduces the class of models to:

$$\mathbf{p} = f(\mathbf{u},\mathbf{v}) \tag{2}$$

However, this still leaves the particular form of the function f unspecified. Now, if we assume that **p** lies in the same space as **u** and **v**, avoiding the issues of dimensionality associated with tensor products, and that f is a linear function, for simplicity, of the cartesian product of **u** and **v**, then we generate a class of *additive* models:

$$\mathbf{p} = \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} \tag{3}$$

where **A** and **B** are matrices which determine the contributions made by **u** and **v** to the product **p**. In contrast, if we assume that f is a linear function of the tensor product of **u** and **v**, then we obtain *multiplicative* models:

$$\mathbf{p} = \mathbf{C}\mathbf{u}\mathbf{v} \tag{4}$$

where **C** is a tensor of rank 3, which projects the tensor product of **u** and **v** onto the space of **p**.

Further constraints can be introduced to reduce the free parameters in these models. So, if we assume that only the ith components of **u** and **v** contribute to the ith component of **p**, that these components are not dependent on i, and that the function is symmetric with regard to the interchange of **u**

[1] A detailed treatment of existing semantic space models is outside the scope of the present paper. We refer the interested reader to Padó and Lapata (2007) for a comprehensive overview.

and **v**, we obtain a simpler instantiation of an additive model:

$$p_i = u_i + v_i \tag{5}$$

Analogously, under the same assumptions, we obtain the following simpler multiplicative model:

$$p_i = u_i \cdot v_i \tag{6}$$

For example, according to (5), the addition of the two vectors representing *horse* and *run* in Figure 1 would yield $\mathbf{horse} + \mathbf{run} = [1\ \ 14\ \ 6\ \ 14\ \ 4]$. Whereas their product, as given by (6), is $\mathbf{horse} \cdot \mathbf{run} = [0\ \ 48\ \ 8\ \ 40\ \ 0]$.

Although the composition model in (5) is commonly used in the literature, from a linguistic perspective, the model in (6) is more appealing. Simply adding the vectors **u** and **v** lumps their contents together rather than allowing the content of one vector to pick out the relevant content of the other. Instead, it could be argued that the contribution of the *i*th component of **u** should be scaled according to its relevance to **v**, and vice versa. In effect, this is what model (6) achieves.

As a result of the assumption of symmetry, both these models are 'bag of words' models and word order insensitive. Relaxing the assumption of symmetry in the case of the simple additive model produces a model which weighs the contribution of the two components differently:

$$p_i = \alpha u_i + \beta v_i \tag{7}$$

This allows additive models to become more syntax aware, since semantically important constituents can participate more actively in the composition. As an example if we set α to 0.4 and β to 0.6, then $\mathbf{horse} = [0\ \ 2.4\ \ 0.8\ \ 4\ \ 1.6]$ and $\mathbf{run} = [0.6\ \ 4.8\ \ 2.4\ \ 2.4\ \ 0]$, and their sum $\mathbf{horse} + \mathbf{run} = [0.6\ \ 5.6\ \ 3.2\ \ 6.4\ \ 1.6]$.

An extreme form of this differential in the contribution of constituents is where one of the vectors, say **u**, contributes nothing at all to the combination:

$$p_i = v_j \tag{8}$$

Admittedly the model in (8) is impoverished and rather simplistic, however it can serve as a simple baseline against which to compare more sophisticated models.

The models considered so far assume that components do not 'interfere' with each other, i.e., that only the *i*th components of **u** and **v** contribute to the *i*th component of **p**. Another class of models can be derived by relaxing this constraint. To give a concrete example, circular convolution is an instance of the general multiplicative model which breaks this constraint by allowing u_j to contribute to p_i:

$$p_i = \sum_j u_j \cdot v_{i-j} \tag{9}$$

It is also possible to re-introduce the dependence on K into the model of vector composition. For additive models, a natural way to achieve this is to include further vectors into the summation. These vectors are not arbitrary and ideally they must exhibit some relation to the words of the construction under consideration. When modeling predicate-argument structures, Kintsch (2001) proposes including one or more distributional neighbors, **n**, of the predicate:

$$\mathbf{p} = \mathbf{u} + \mathbf{v} + \sum \mathbf{n} \tag{10}$$

Note that considerable latitude is allowed in selecting the appropriate neighbors. Kintsch (2001) considers only the m most similar neighbors to the predicate, from which he subsequently selects k, those most similar to its argument. So, if in the composition of *horse* with *run*, the chosen neighbor is *ride*, $\mathbf{ride} = [2\ \ 15\ \ 7\ \ 9\ \ 1]$, then this produces the representation $\mathbf{horse} + \mathbf{run} + \mathbf{ride} = [3\ \ 29\ \ 13\ \ 23\ \ 5]$. In contrast to the simple additive model, this extended model is sensitive to syntactic structure, since **n** is chosen from among the neighbors of the predicate, distinguishing it from the argument.

Although we have presented multiplicative and additive models separately, there is nothing inherent in our formulation that disallows their combination. The proposal is not merely notational. One potential drawback of multiplicative models is the effect of components with value zero. Since the product of zero with any number is itself zero, the presence of zeroes in either of the vectors leads to information being essentially thrown away. Combining the multiplicative model with an additive model, which does not suffer from this problem, could mitigate this problem:

$$p_i = \alpha u_i + \beta v_i + \gamma u_i v_i \tag{11}$$

where α, β, and γ are weighting constants.

4 Evaluation Set-up

We evaluated the models presented in Section 3 on a sentence similarity task initially proposed by Kintsch (2001). In his study, Kintsch builds a model of how a verb's meaning is modified in the context of its subject. He argues that the subjects of *ran* in *The color ran* and *The horse ran* select different senses of *ran*. This change in the verb's sense is equated to a shift in its position in semantic space. To quantify this shift, Kintsch proposes measuring similarity relative to other verbs acting as landmarks, for example *gallop* and *dissolve*. The idea here is that an appropriate composition model when applied to *horse* and *ran* will yield a vector closer to the landmark *gallop* than *dissolve*. Conversely, when *color* is combined with *ran*, the resulting vector will be closer to *dissolve* than *gallop*.

Focusing on a single compositional structure, namely intransitive verbs and their subjects, is a good point of departure for studying vector combination. Any adequate model of composition must be able to represent argument-verb meaning. Moreover by using a minimal structure we factor out inessential degrees of freedom and are able to assess the merits of different models on an equal footing. Unfortunately, Kintsch (2001) demonstrates how his own composition algorithm works intuitively on a few hand selected examples but does not provide a comprehensive test set. In order to establish an independent measure of sentence similarity, we assembled a set of experimental materials and elicited similarity ratings from human subjects. In the following we describe our data collection procedure and give details on how our composition models were constructed and evaluated.

Materials and Design Our materials consisted of sentences with an an intransitive verb and its subject. We first compiled a list of intransitive verbs from CELEX[2]. All occurrences of these verbs with a subject noun were next extracted from a RASP parsed (Briscoe and Carroll, 2002) version of the British National Corpus (BNC). Verbs and nouns that were attested less than fifty times in the BNC were removed as they would result in unreliable vectors. Each reference subject-verb tuple (e.g., *horse ran*) was paired with two landmarks, each a synonym of the verb. The landmarks were chosen so as to represent distinct verb senses, one compatible with the reference (e.g., *horse galloped*) and one incompatible (e.g., *horse dissolved*). Landmarks were taken from WordNet (Fellbaum, 1998). Specifically, they belonged to different synsets and were maximally dissimilar as measured by the Jiang and Conrath (1997) measure.[3]

Our initial set of candidate materials consisted of 20 verbs, each paired with 10 nouns, and 2 landmarks (400 pairs of sentences in total). These were further pretested to allow the selection of a subset of items showing clear variations in sense as we wanted to have a balanced set of similar and dissimilar sentences. In the pretest, subjects saw a reference sentence containing a subject-verb tuple and its landmarks and were asked to choose which landmark was most similar to the reference or neither. Our items were converted into simple sentences (all in past tense) by adding articles where appropriate. The stimuli were administered to four separate groups; each group saw one set of 100 sentences. The pretest was completed by 53 participants.

For each reference verb, the subjects' responses were entered into a contingency table, whose rows corresponded to nouns and columns to each possible answer (i.e., one of the two landmarks). Each cell recorded the number of times our subjects selected the landmark as compatible with the noun or not. We used Fisher's exact test to determine which verbs and nouns showed the greatest variation in landmark preference and items with p-values greater than 0.001 were discarded. This yielded a reduced set of experimental items (120 in total) consisting of 15 reference verbs, each with 4 nouns, and 2 landmarks.

Procedure and Subjects Participants first saw a set of instructions that explained the sentence similarity task and provided several examples. Then the experimental items were presented; each contained two sentences, one with the reference verb and one with its landmark. Examples of our items are given in Table 1. Here, *burn* is a high similarity landmark (High) for the reference *The fire glowed*, whereas *beam* is a low similarity landmark (Low). The opposite is the case for the reference *The face*

[2] http://www.ru.nl/celex/

[3] We assessed a wide range of semantic similarity measures using the WordNet similarity package (Pedersen et al., 2004). Most of them yielded similar results. We selected Jiang and Conrath's measure since it has been shown to perform consistently well across several cognitive and NLP tasks (Budanitsky and Hirst, 2001).

Noun	Reference	High	Low
The fire	glowed	burned	beamed
The face	glowed	beamed	burned
The child	strayed	roamed	digressed
The discussion	strayed	digressed	roamed
The sales	slumped	declined	slouched
The shoulders	slumped	slouched	declined

Table 1: Example Stimuli with High and Low similarity landmarks

glowed. Sentence pairs were presented serially in random order. Participants were asked to rate how similar the two sentences were on a scale of one to seven. The study was conducted remotely over the Internet using Webexp[4], a software package designed for conducting psycholinguistic studies over the web. 49 unpaid volunteers completed the experiment, all native speakers of English.

Analysis of Similarity Ratings The reliability of the collected judgments is important for our evaluation experiments; we therefore performed several tests to validate the quality of the ratings. First, we examined whether participants gave high ratings to high similarity sentence pairs and low ratings to low similarity ones. Figure 2 presents a box-and-whisker plot of the distribution of the ratings. As we can see sentences with high similarity landmarks are perceived as more similar to the reference sentence. A Wilcoxon rank sum test confirmed that the difference is statistically significant ($p < 0.01$). We also measured how well humans agree in their ratings. We employed leave-one-out resampling (Weiss and Kulikowski, 1991), by correlating the data obtained from each participant with the ratings obtained from all other participants. We used Spearman's ρ, a non parametric correlation coefficient, to avoid making any assumptions about the distribution of the similarity ratings. The average inter-subject agreement[5] was $\rho = 0.40$. We believe that this level of agreement is satisfactory given that naive subjects are asked to provide judgments on fine-grained semantic distinctions (see Table 1). More evidence that this is not an easy task comes from Figure 2 where we observe some overlap in the ratings for High and Low similarity items.

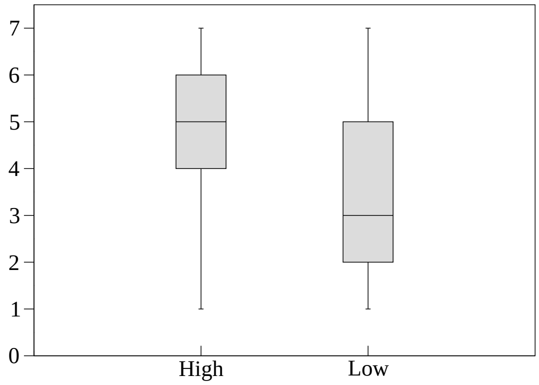

Figure 2: Distribution of elicited ratings for High and Low similarity items

不及物动词在搭配不同的主语时动词的含义发生变化

The fire glowed(reference)
The fire burn(landmark)

Model Parameters Irrespectively of their form, all composition models discussed here are based on a semantic space for representing the meanings of individual words. The semantic space we used in our experiments was built on a lemmatised version of the BNC. Following previous work (Bullinaria and Levy, 2007), we optimized its parameters on a word-based semantic similarity task. The task involves examining the degree of linear relationship between the human judgments for two individual words and vector-based similarity values. We experimented with a variety of dimensions (ranging from 50 to 500,000), vector component definitions (e.g., pointwise mutual information or log likelihood ratio) and similarity measures (e.g., cosine or confusion probability). We used WordSim353, a benchmark dataset (Finkelstein et al., 2002), consisting of relatedness judgments (on a scale of 0 to 10) for 353 word pairs.

We obtained best results with a model using a context window of five words on either side of the target word, the cosine measure, and 2,000 vector components. The latter were the most common context words (excluding a stop list of function words). These components were set to the ratio of the probability of the context word given the target word to the probability of the context word overall. This configuration gave high correlations with the WordSim353 similarity judgments using the cosine measure. In addition, Bullinaria and Levy (2007) found that these parameters perform well on a number of other tasks such as the synonymy task from the *Test of English as a Foreign Language* (TOEFL).

模型空间元素的含义

Our composition models have no additional pa-

[4] http://www.webexp.info/

[5] Note that Spearman's rho tends to yield lower coefficients compared to parametric alternatives such as Pearson's r.

rameters beyond the semantic space just described, with three exceptions. First, the additive model in (7) weighs differentially the contribution of the two constituents. In our case, these are the subject noun and the intransitive verb. To this end, we optimized the weights on a small held-out set. Specifically, we considered eleven models, varying in their weightings, in steps of 10%, from 100% noun through 50% of both verb and noun to 100% verb. For the best performing model the weight for the verb was 80% and for the noun 20%. Secondly, we optimized the weightings in the combined model (11) with a similar grid search over its three parameters. This yielded a weighted sum consisting of 95% verb, 0% noun and 5% of their multiplicative combination. Finally, Kintsch's (2001) additive model has two extra parameters. The m neighbors most similar to the predicate, and the k of m neighbors closest to its argument. In our experiments we selected parameters that Kintsch reports as optimal. Specifically, m was set to 20 and m to 1.

Evaluation Methodology We evaluated the proposed composition models in two ways. First, we used the models to estimate the cosine similarity between the reference sentence and its landmarks. We expect better models to yield a pattern of similarity scores like those observed in the human ratings (see Figure 2). A more scrupulous evaluation requires directly correlating all the individual participants' similarity judgments with those of the models.[6] We used Spearman's ρ for our correlation analyses. Again, better models should correlate better with the experimental data. We assume that the inter-subject agreement can serve as an upper bound for comparing the fit of our models against the human judgments.

5 Results

Our experiments assessed the performance of seven composition models. These included three additive models, i.e., simple addition (equation (5), Add), weighted addition (equation (7), WeightAdd), and Kintsch's (2001) model (equation (10), Kintsch), a multiplicative model (equation (6), Multiply), and also a model which combines multiplication with addition (equation (11), Combined). As a baseline we simply estimated the similarity between the reference verb and its landmarks without taking the subject noun into account (equation (8), NonComp). Table 2 shows the average model ratings for High and Low similarity items. For comparison, we also show the human ratings for these items (UpperBound). Here, we are interested in relative differences, since the two types of ratings correspond to different scales. Model similarities have been estimated using cosine which ranges from 0 to 1, whereas our subjects rated the sentences on a scale from 1 to 7.

Model	High	Low	ρ
NonComp	0.27	0.26	0.08**
Add	0.59	0.59	0.04*
WeightAdd	0.35	0.34	0.09**
Kintsch	0.47	0.45	0.09**
Multiply	0.42	0.28	0.17**
Combined	0.38	0.28	0.19**
UpperBound	4.94	3.25	0.40**

Table 2: Model means for High and Low similarity items and correlation coefficients with human judgments (*: $p < 0.05$, **: $p < 0.01$)

The simple additive model fails to distinguish between High and Low Similarity items. We observe a similar pattern for the non compositional baseline model, the weighted additive model and Kintsch (2001). The multiplicative and combined models yield means closer to the human ratings. The difference between High and Low similarity values estimated by these models are statistically significant ($p < 0.01$ using the Wilcoxon rank sum test). Figure 3 shows the distribution of estimated similarities under the multiplicative model.

The results of our correlation analysis are also given in Table 2. As can be seen, all models are significantly correlated with the human ratings. In order to establish which ones fit our data better, we examined whether the correlation coefficients achieved differ significantly using a t-test (Cohen and Cohen, 1983). The lowest correlation ($\rho = 0.04$) is observed for the simple additive model which is not significantly different from the non-compositional baseline model. The weighted additive model ($\rho = 0.09$) is not significantly different from the baseline either or Kintsch (2001) ($\rho = 0.09$). Given that the basis

[6] We avoided correlating the model predictions with averaged participant judgments as this is inappropriate given the ordinal nature of the scale of these judgments and also leads to a dependence between the number of participants and the magnitude of the correlation coefficient.

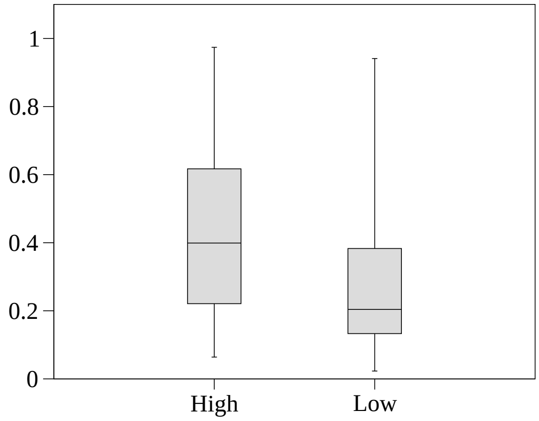

Figure 3: Distribution of predicted similarities for the vector multiplication model on High and Low similarity items

of Kintsch's model is the summation of the verb, a neighbor close to the verb and the noun, it is not surprising that it produces results similar to a summation which weights the verb more heavily than the noun. The multiplicative model yields a better fit with the experimental data, $\rho = 0.17$. The combined model is best overall with $\rho = 0.19$. However, the difference between the two models is not statistically significant. Also note that in contrast to the combined model, the multiplicative model does not have any free parameters and hence does not require optimization for this particular task.

6 Discussion

In this paper we presented a general framework for vector-based semantic composition. We formulated composition as a function of two vectors and introduced several models based on addition and multiplication. Despite the popularity of additive models, our experimental results showed the superiority of models utilizing multiplicative combinations, at least for the sentence similarity task attempted here. We conjecture that the additive models are not sensitive to the fine-grained meaning distinctions involved in our materials. Previous applications of vector addition to document indexing (Deerwester et al., 1990) or essay grading (Landauer et al., 1997) were more concerned with modeling the gist of a document rather than the meaning of its sentences. Importantly, additive models capture composition by considering *all* vector components representing the meaning of the verb and its subject, whereas multiplicative models consider a subset, namely non-zero components. The resulting vector is sparser but expresses more succinctly the meaning of the predicate-argument structure, and thus allows semantic similarity to be modelled more accurately.

Further research is needed to gain a deeper understanding of vector composition, both in terms of modeling a wider range of structures (e.g., adjective-noun, noun-noun) and also in terms of exploring the space of models more fully. We anticipate that more substantial correlations can be achieved by implementing more sophisticated models from within the framework outlined here. In particular, the general class of multiplicative models (see equation (4)) appears to be a fruitful area to explore. Future directions include constraining the number of free parameters in linguistically plausible ways and scaling to larger datasets.

The applications of the framework discussed here are many and varied both for cognitive science and NLP. We intend to assess the potential of our composition models on context sensitive semantic priming (Till et al., 1988) and inductive inference (Heit and Rubinstein, 1994). NLP tasks that could benefit from composition models include paraphrase identification and context-dependent language modeling (Coccaro and Jurafsky, 1998).

References

E. Briscoe, J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, 1499–1504, Las Palmas, Canary Islands.

A. Budanitsky, G. Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of ACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA.

J. Bullinaria, J. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

F. Choi, P. Wiemer-Hastings, J. Moore. 2001. Latent semantic analysis for text segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 109–117, Pittsburgh, PA.

N. Coccaro, D. Jurafsky. 1998. Towards better integration of semantic predictors in statistical language modeling. In *Proceedings of the 5th International Conference on Spoken Language Processsing*, Sydney, Australia.

J. Cohen, P. Cohen. 1983. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, R. A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

G. Denhire, B. Lemaire. 2004. A computational model of children's semantic memory. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, 297–302, Chicago, IL.

C. Fellbaum, ed. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, E. Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

J. Fodor, Z. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28:3–71.

P. W. Foltz, W. Kintsch, T. K. Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Process*, 15:285–307.

S. Frank, M. Koppen, L. Noordman, W. Vonk. 2007. World knowledge in computational models of discourse comprehension. *Discourse Processes*. In press.

G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

Z. Harris. 1968. *Mathematical Structures of Language*. Wiley, New York.

E. Heit, J. Rubinstein. 1994. Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:411–422.

J. J. Jiang, D. W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*, Taiwan.

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

T. K. Landauer, S. T. 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.

T. K. Landauer, D. Laham, B. Rehder, M. E. Schreiner. 1997. How well can passage meaning be derived without using word order: A comparison of latent semantic analysis and humans. In *Proceedings of 19th Annual Conference of the Cognitive Science Society*, 412–417, Stanford, CA.

K. Lund, C. Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28:203–208.

D. McCarthy, R. Koeling, J. Weeds, J. Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 280–287, Barcelona, Spain.

S. McDonald. 2000. *Environmental Determinants of Lexical Processing Effort.* Ph.D. thesis, University of Edinburgh.

R. Montague. 1974. English as a formal language. In R. Montague, ed., *Formal Philosophy*. Yale University Press, New Haven, CT.

H. Neville, J. L. Nichol, A. Barss, K. I. Forster, M. F. Garrett. 1991. Syntactically based sentence prosessing classes: evidence form event-related brain potentials. *Journal of Congitive Neuroscience*, 3:151–165.

S. Padó, M. Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

T. Pedersen, S. Patwardhan, J. Michelizzi. 2004. WordNet::similarity - measuring the relatedness of concepts. In *Proceedings of the 5th Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 38–41, Boston, MA.

T. A. Plate. 1991. Holographic reduced representations: Convolution algebra for compositional distributed representations. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, 30–35, Sydney, Australia.

G. Salton, A. Wong, C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

P. Schone, D. Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 100–108, Pittsburgh, PA.

H. Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

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

R. E. Till, E. F. Mross, W. Kintsch. 1988. Time course of priming for associate and inference words in discourse context. *Memory and Cognition*, 16:283–299.

S. M. Weiss, C. A. Kulikowski. 1991. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, San Mateo, CA.

R. F. West, K. E. Stanovich. 1986. Robust effects of syntactic structure on visual word processing. *Journal of Memory and Cognition*, 14:104–112.