Compositional Expectation: A Purely Distributional Model of Compositional Semantics

Justin Washtell
University of Leeds, UK
washtell@comp.leeds.ac.uk

The past year has witnessed a surge of interest in the issue of compositional semantics: modelling the meaning of complex phrases. To date, distributional approaches have successfully dealt only with the meaning of individual words in context. Recent attempts to address the more general case of compositional meaning have tended to focus either on mathematical models, which have yet to be demonstrated useful in a linguistic setting, or on syntactically-motivated approaches which do not yet permit application to unconstrained text. We present a purely distributional compositional model, based on the simple addition of expectation vectors. Expectation vectors (Washtell, 2010) are particularly appealing from a compositional standpoint as they are naturally sensitive to word-order alterations whilst being insensitive to the substitution of distributionally similar words. We explore the properties of these and two baseline models using datasets based upon human judgements of phrasal similarity. Whilst far from solving the problem of compositionality, our findings raise interesting questions and provide some useful ideas and benchmarks for those tackling this very current problem.

1. Introduction and motivation

The distributional hypothesis has enjoyed great success over the past decade in the field of empirical lexical semantics, with distributional models performing competitively in tasks which would once have been considered the province of knowledge-driven systems. Many of these successes have been based upon geometric/vector representations and have dealt with the classification of individual words in generalised or specific contexts, in which the unstructured content of those contexts has proven sufficient to enable useful inferences regarding the semantics of the word in question. Such purely distributional models of semantics are particularly attractive from a cognitive perspective as they presuppose little language-specific knowledge, and thus can also be construed as models of language acquisition[1]. This also represents a practical benefit for applications-oriented research as such models can be applied to various languages and registers with little modification.

However if distributional models, or indeed semantic models in general, are to describe language adequately and continue their success, then the issue of compositional meaning remains to be addressed (Pustejovsky, 1995). Baroni (2010) observes that a *"long tradition of scholars unsympathetic to statistical* [i.e. distributional] *approaches to language have argued that they are doomed to fail because they cannot capture compositionality"*. As limits are being reached regarding what can be accomplished with existing (i.e. non-compositional) models, and the growing volume of un-annotated digital content continues to motivate a search for ever more sophisticated ways of making sense of it, the importance of Baroni's scholars' challenge is beginning to become acutely felt in the research community. 2010 saw the introduction of the ESSLLI workshop on Compositionality and Distributional Semantic Models (DistComp), conceived specifically to address this problem. At the same time, over half of the papers at the ACL 2010 workshop on Geometrical Models of Natural Language Semantics (GEMS'10) dealt explicitly with the issue of compositionality (in contrast with just a single paper the previous year).

In this paper we present a purely distributional vector model of semantic compositionality. The model is based on work inspired by principles in semiotics (Washtell, 2010), a field which traditionally has philosophical and sociolinguistic leanings. However, the main idea exploited – that of the central role of expectation in meaning – has begun to receive attention elsewhere in the lexical

[1] In reality nearly all present models incorporate *some* language or task specific knowledge; the most pervasive is the need to define what constitutes a lexical unit (token), without which distributional analysis is problematic. The fact that distributional regularities can be observed on many scales suggest that this is an uneasy solution.

semantics literature (Erk & Padó, 2008, 2009; Yuret, 2007, 2010) and has support in both psycholinguistics and information theory (Attneave, 1959; Brouwer *et al*, 2010; Mitchell *et al*, 2010).

2. Background

Frege's *principle of compositionality* asserts that the meaning of a complex expression is determined by the *meanings of its parts* and the *way in which those parts are combined*. This seems somewhat at odds with the distributional hypothesis; whereas the latter links the meaning of an utterance to its external context, the former focuses on the internal; together these create an apparently circular delegation of responsibility concerning the residence of meaning[2]. In empirical lexical semantics, it is the distributional hypothesis which has received the most attention. This accords with the fact that most attempts to model meaning have focused on atomic units (i.e. words or lemmas), this either being seen as a necessary step towards conquering compositional meaning, or having provided sufficient gains in its own right to distract from what is generally considered a harder problem.

Perhaps one of the simplest geometric model of phrase meaning found in the literature is the bag-of-vectors (more generally the "bag-of-words") in which a vector model of word meaning is elevated to the phrase or document level by summing (or alternatively performing component-wise multiplication[3]) of word vectors (Schütze, 1998). This has the advantage that it is immediately applicable to any model in which word meaning can be expressed as a vector. The principle limitation of course is that this takes no account of word order. Therefore, while it has proven effective in the context of document retrieval and coarse-grained classification tasks, it is generally considered insufficient for fine-grained semantic tasks such as lexical entailment and question answering in which structure tends to play a dominant role. One major line of investigation therefore has been into vector models which are capable of encoding word order information. Vector representations, despite their convenience, seem to present something of a hurdle in this respect. One obvious approach is to use a non-commutative vector operator, such as the tensor product (Smolensky, 1990). Alas, as the product of two tensors results in a tensor of higher order (e.g. a matrix from a vector), the dimensionality of the representation increases exponentially with each term added. This obviously presents a problem for the meaningful comparison of phrases of different lengths, not to mention scalability.

Circular convolution has been proposed as a non-commutative vector operator which does not suffer from the problem of dimensionality explosion (Plate, 1995; Jones & Mewhort, 2007). Sahlgren *et al* (2008) describe an alternative vector-based approach to combining words-in-context with word-order information which does not focus on a specific operator. Rather, a vector for a given word is "contextualized" by merging it with permuted forms of the vectors representing its neighbouring words. After this, simple vector addition is used[4]. The number of times a vector is permuted depends on its distance from the word of interest. Owing to this explicit dependence on word positions, the method is only proposed as a way of comparing individual words in context, not arbitrary passages (i.e. such that context words are imbued with meaningful positions relative to the headword slot). Both convolution and permutation seem like somewhat heavy-handed ways of encoding structure for semantic applications: while structures having similar words at identical positions may be compositionally similar under these approaches (depending on how those words' vectors are formed), structures having similar or even identical words at slightly different positions will not be, as any similarities will have been obfuscated by the permutation or convolution process. Convolutions in particular were designed to operate upon periodic functions and time-series data; the manner in which they discard information in a linguistic context seems rather arbitrary. Unsurprisingly, Mitchell & Lapata (2010) find compositional models built upon convolution to perform very poorly.

Rudolph & Giesbrecht (2010) have proposed square matrices as an alternative to vectors for building compositional models. Standard matrix multiplication is both non-commutative and will take

[2] Some citations of Frege's principle bring the distributional and compositional hypotheses into even starker conflict, observing that the meaning of a part is the *contribution* it makes to the phrase *to which it belongs*.

[3] In vectors comprising positive non-zero elements, these operators are equivalent: adding components such as PMI which incorporates a log function, is equivalent to multiplying them in the absence of the log. The preferred function is therefore dependent upon the nature of the vector components (see Mitchell & Lapata, 2010).

[4] Although Sahlgren *et al*'s model is not cast in terms of operators, it can be thought of as involving a non-commutative operator which entails a permutation and an addition, so making it comparable to a convolution.

two square matrices and produce another matrix of the same dimensions. The authors show that such a matrix representation is able to subsume various existing vector models (e.g. circular convolution) by varying the manner in which vectors are encoded into matrices. Nonetheless, it is not yet clear how this approach can be used to transcend the existing limitations of these models.

Specifically, while the various approaches described attempt to address the issue of differing linguistic forms requiring different representations, it is not clear how any of them might usefully capture structural synonymy, in which markedly different forms have very similar meanings (e.g. *[noun1][passive verb] [noun2]* versus *[noun2][active verb][noun1]*), or differentiate either case from those in which word order is relatively unimportant (e.g. *"they researched it thoroughly"* versus *"they thoroughly researched it"*). Also, while going to lengths to compose words in a way that is non-commutative, these methods largely assume *associativity*. While in some cases this may seem inconsequential: *dogs (chase cats)* ≈ *(dogs chase) cats*. In other, especially less compositional (more lexicalised) cases, it seems unsatisfactory: *(new york) skyline* ≠ *new (york skyline)*. Arguably what is needed are models which contend intelligently with varying degrees of compositionality. This would seem to favour something more sophisticated than can be captured by a mathematical operator.

The limitations of mathematical approaches to "compositionalising" distributional models are perhaps one reason why many researchers have eschewed a purely distributional approach in favour of more linguistically informed models, incorporating notions of word or relationship *type* (Padó & Lapata, 2007; Clark *et al* 2008). A full review of these is beyond the scope of this paper, suffice to say that some of these models have demonstrated promise under restricted experimental conditions. Of particular relevance to us are the approaches taken by Erk & Padó (2008, 2009) and Thater *et al* (2010) in which a component word is represented by combining a vector describing its type with one describing the selectional preferences (or expectations) of one of its dependent terms, in parsed text. In this way, words provide context for each other, and the two newly contextualized words define the meaning of the whole: an idea known as *co-compositionality* (Putsejovsky, 1995; Baroni, 2010; Gamallo *et al,* 2004). Gayral *et al* (2000) argue that this type of compositionality alone (see also Kintsch & Mangalath, 2010) is insufficient, and that compositional meaning is dependent on features which go beyond immediate arguments. Erk & Padó acknowledge that the generalisation of their approach to multi-word contexts is an open problem. Baroni (2010) speculates that some form of "recursive" compositionality to this end ought to be achievable, providing that the principles governing when and how words influence each other's meaning can be resolved.

3. Expectation vectors

Expectation Vectors were introduced by Washtell (2010) as an intuitive way of modelling the meaning of a word-in-context. An expectation vector for any context or word-in-context can be formed by applying a predictive language model to that context and generating a distribution over word types in the lexicon which reflects the likelihood of each word occurring in the headword-slot. This distribution can then be treated as a vector, and similarity comparisons performed using standard wordspace techniques. The attraction of these vectors lies in perhaps three key features. The first is the intuitive way in which word meaning is modelled: not in terms of a set of context features, but rather the set of words which can be substituted in a given context. It is reasoned that abstracting away from context features in this manner allows for similarity metrics which more directly capture phenomena such as polysemy and synonymy. Secondly, this separation allows for the leveraging of arbitrarily sophisticated language models, such as are able to capture complex interdependencies between words in use and incorporate broader contextual information without the need to complicate the resultant vector space. Thirdly, this can result in markedly denser vectors than using surface features directly. Washtell (2010) and Yuret (2007, 2010) both found approaches based on expectation to perform well in word sense disambiguation tasks, hypothesizing that data-density plays a key role in this setting.

Another, as yet unexplored, advantage of methods based on expectation is that they seem to lend themselves particularly well to modelling compositional meaning. It is this benefit which is the focus of the present work. The key observation is that, providing a non-trivial language model is employed, expectation vectors are naturally word-order dependent. Thus, unlike previously proposed approaches such as that of Sahlgren (2008), which hinge upon the post-hoc manipulation or contextualization of word-type vectors, a word-instance and its context are much more fundamentally intertwined.

In this work we compute a *compositional* vector for an arbitrary fragment of text by generating an expectation vector **e** for every word position in that text (using the remaining words as context in each case) and then simply summing. For generating expectations, we take the approach described in Washtell (2010), in which a structural similarity metric compares the candidate context **c** to the context of every word position *o* in a large corpus *O*:

$$\mathbf{e}_j = P(j|c) \sim \max_{o_i^k \in O_i} \text{sim}(o_j^k, c)$$

Where O_j represents the set of contexts for word type *j* in the corpus, and *k* is a specific instance of *j*. The maximum similarity score across all instances of a word type in the corpus therefore forms that word type's corresponding vector component. In this way the similarity metric constitutes a general language model and, along with a raw corpus, a specific language model. This is a computationally expensive approach. Washtell (2010) presents a cumbersome similarity metric based on distance ratios. Here we take a simpler and more efficient approach. First, we form context vectors by summing the negative exponents of each word type's occurrence positions relative to the context head. That is to say that in the context *"the cat sat on the _"*, *the* will have a value of $b^{-1} + b^{-5}$, *on* will have a value of b^{-2}, and so on, with the base *b* constituting a distance falloff parameter. Similarity is then computed by simply taking the square of the dot product of the two vectors. When $b>1$, this product will never exceed a constant value, irrespective of the context size, thus avoiding the need for normalization. As well as giving an intuitive measure of similarity (effectively computing a "structural" correlation), this approach has the advantage that similarities can be calculated incrementally as we pass through the corpus. The complexity of calculating an expectation vector is therefore more-or-less linear with the size of the corpus, irrespective of the size of the supplied context. Further optimizations can be made if we observe that, for $b>=2$, matching a single pair of words at a given distance from the head always results in a greater similarity than matching any number of words further away.

For the evaluations performed herein we use the British National Corpus and a value of $b=2$. This was found to generate subjectively coherent and cohesive text when recursively extending a context by selecting one of its higher-ranking expectations: a promising trait for the generation of meaningful expectation vectors. The remaining details differentiating our approach from that in Washtell (2010) lie in the handling of the vector components. First, the vectors are normalized so that their components sum to one, giving a pseudo-probabilistic distribution. We then divide each component by its respective word type's prior probability (i.e. its frequency in the corpus) to give a set of probability ratios. Practically speaking these steps prevent function words, and vague expectations which comprise many equally likely words, from routinely dominating our compositional vectors.

4. Evaluation

Distributional wordspace models are often evaluated on their ability to capture meaning by comparing the predictions of their similarity metrics with datasets encapsulating human judgements of word similarity. Achananuparp *et al* (2008) and Mitchell & Lapata (2010) have extended this philosophy to evaluating a variety of phrasal similarity measures. The former rely on human-annotated paraphrase and entailment datasets. Arguably these datasets concern themselves with a much narrower notion of similarity than do word-oriented studies: that of a kind of logical or truth-conditional equivalence. This binary concept does not sit particularly well with vector models, in which meaning is considered to occupying a continuum. Nor does it allow for, say, analogous meanings, or statements of fact re-expressed as questions or opinions. As we cannot say *a priori* that any particular type of relationship plays a dominant role in human intuitions of meaning, it seems unreasonable to exclude any from our investigations; from a application-oriented perspective, if we can first establish *what* it is that our compositional approaches capture, then we will be better placed to pursue specific types of meaning.

Mitchell & Lapata (2010) build a dataset from the ground up which is arguably better suited to this task (hereafter the "M&L dataset"), using phrases extracted from the BNC with the aid of heuristics conceived to capture a range and variety of semantic similarities. We adopt their dataset here, as we believe it provides a sound starting point for evaluating compositional models. We then go on to describe a complementary evaluation in which we attempt to address some of the weaknesses inherent in the M&L dataset in order to provide a more holistic picture. We compare three similarity

measures across these evaluations: our proposed expectation-oriented approach and two baselines, each outlined below. In keeping with our purely distributional interests, no language-specific pre-processing steps such as lemmatisation, POS-tagging or parsing were used with any of the measures in either of the evaluations.

Bigram overlap (BIGRAM) is simply the total number of character bigrams that two phrases have in common, normalized by the total number of bigrams they collectively possess (i.e. the Jaccard coefficient). Identical strings achieve a similarity score of 1, with less similar strings having scores that tend towards zero. The main advantages of this approach in the settings herein is that it is forgiving of small changes in word or clause order, and in the inflected forms of words, which in many cases may not significantly affect meaning. By the same token however, it is insensitive to significant shifts in meaning which can sometimes be induced in this way (for example, by the swapping of subject and object). The other main disadvantage of the character bigram model is that, being a simple string similarity measure, it is fundamentally incapable of acknowledging similarities in meaning between completely different forms (i.e. synonymy).

Bag-of-vectors (VECTORBAG) entails summing co-occurrence vectors for the component words of a phrase, where those vectors are derived from a large corpus containing examples of the words in context. Summed vectors are then compared using cosine similarity, which ignores the vector size (effectively factoring out phrase length), focusing instead on the relative balance of components present. This is comparable to the higher-order approaches taken by Schütze (1998) and Landauer & Dumais (1997). Note however that we use a distance-based association measure *co-dispesion* to construct word vectors (Washtell & Markert, 2010). As well as avoiding the thorny issue of window-size and arguably providing a better exploitation of the data in general, this provides a more meaningful baseline for our expectation model which uses a distance-based language model. The principle advantage of working with co-occurrence vectors is that distributionally similar words become comparable by virtue of their vectors being similar. As word-type vectors are the centroids of all occurrences in a corpus, senses are conflated, so synonymy is not modelled particularly cleanly, and polysemy arguably not at all. However the thinking is that when combined in a phrase, the common semantic components of the words dominate, with incidental senses being relegated to some acceptable level of noise. The major disadvantage with such a "bagged" approach is that word order is entirely discarded; whereas under the bigram model switching subject and object would at least incur a small penalty, here the two resultant phrases appear entirely equivalent.

As with VECTORBAG, in compositional expectation (COMPEXP) phrase vectors are formed by summing the vectors of their component words, and then compared using cosine similarity. Rather than the component vectors being based upon word-types however, we use expectation vectors (see section 3) which are unique to the phrasal context in which each word occurs.

4.1. Evaluation 1: Simple Phrase Similarity

Our first method of evaluation is against the M&L phrase similarity dataset (see section 4). This consists of around 200 short phrase pairs rated by human subjects on a scale of 1-7 for their semantic similarity. Each phrase is comprised of two words in the form verb-object, noun-noun, or adjective-noun, extracted from the BNC. The authors applied quite sophisticated heuristics based on phrase frequency and WordNet word similarity (Lesk, 1986) in an attempt to produce a set which exhibits an even spread of subjective similarities, from near-synonymy to near-total unrelatedness. Their analysis of human ratings confirmed that they were reasonably successful in this.

Table 1 shows the performance of the models upon the M&L dataset, in terms of Spearman's rank correlation. Two additional columns are included for reference: the inter-annotator agreement reported by M&L (which in this experiment serves as an upper-bound), calculated using leave-one-out sampling, and the results from the best-performing model reported by M&L for each phrase class. Our additive distance-based model performs fairly competitively on this task, and is superior to all methods on verb-object combinations. Interestingly, compositional expectation fairs relatively poorly, turning in a respectable performance only for noun-noun combinations and performing particularly poorly on verb-object combinations. This last observation is at odds with the surprisingly good performance on verb sense disambiguation previously observed using expectation vectors (Washtell, 2010), leading us to speculate whether this was rather a symptom of the distance-based approach used

in that work (although it should be noted that there are many confounding differences separating the tasks and models in these works). Unsurprisingly, the bigram measure performs very poorly across the board. When interpreting these figures it is worth bearing in mind that, differences in our approaches to composition aside, unlike Mitchell & Lapata we are operating on unlemmatized data.

	Human	M&L BEST (various)	BIGRAM	VECTORBAG	COMPEXP
ADJ-NOUN	0.52	**0.46** (multiplicative)	0.2	0.27	0.28
NOUN-NOUN	0.49	**0.49** (multiplicative)	-0.11	0.47	0.41
VERB-OBJ	0.55	0.41 ("dilated" LDA)	0.11	**0.45**	0.2

Table 1: Spearman's rank correlation between human and computational similarity ratings for M&L dataset.

While the form of the M&L dataset makes it suitable for use as a Gold Standard, it does come with certain limitations. Most notably, the phrases comprise only two words. While this is a logical starting point for assessing compositional models, it gives little scope for testing the ability to capture structural aspects of composition (as M&L anyway restrict phrase pairs to identically structured phrase types, this point is all but moot). Related to this is the fact that while the heuristics applied in generating the M&L dataset attempt to generate superficially different yet synonymous phrases (e.g *"reduce amount", "cut cost"*), there are very few cases of polysemy or superficial similarity (e.g. *"stout Russian", "Russian stout"* or *"arresting music", "arresting criminals"*) which is an important confounding issue for compositional models. In the next section we outline a complementary evaluation approach with which we attempt to address some of these issues.

4.2. Evaluation 2: Unconstrained Phrase Similarity

A restricted register of about 300 noun, verb and adjective lemmas was selected with the aid of Wordnet and BNC frequency information. Sentences were then automatically selected from the BNC with the constraint that each sentence was at least 3 words in length and a certain minimum proportion of its lemmas belonged to the restricted register. This minimum proportion was tweaked such that the entire BNC generated approximately 1000 qualifying sentences. The aim was to produce a manageably sized collection of real-world phrases wherein a range of similarities and similarity types (both semantic and superficial) existed between a proportion of the phrases. In selecting the register, the purpose was therefore to find a compact set of words which exhibited both a high degree of ambiguity (polysemy) and interchangeability (synonymy). Because words satisfying the former requirement tend to be very frequent (e.g. the auxiliary verbs), while those satisfying the latter tend to be very rare, this task was difficult. An additional complication was that the types of phrase selected from the corpus were found to be highly sensitive to the specific words in the register, with certain words resulting in a disproportionate contingent of highly synonymous idiomatic phrases being selected (*"let's take the following", "consider the following", "look at the following"* etc), which was considered undesirable. In the end a lot of judgement was exercised in selecting the register.

For each phrase in the dataset, the two most similar candidate phrases also in the dataset were identified according to each of our three similarity measures. An additional two candidate phrases were selected at random to act as a control. This resulted in at most eight candidate sentences for each source sentence, and less where different methods selected the same phrases. Agreement between BIGRAM and VECTORBAG was 19.7% (which is both surprising and reassuring, considering the size of the dataset and how different these approaches are). Agreement between these and the novel COMPEXP method was markedly less, at 12.7% and 9.1% respectively. Agreement between the random control and each of the methods was in keeping with chance (<0.4%). To aid annotation, further steps were taken to reduce the size of the dataset and increase the proportion of subjectively similar phrases expressed. For each method, the top candidate phrase attributed to each source phrase was ranked amongst those attributed to all source phrases (according to the actual similarity score attributed). The source phrases were then ordered according to the minimum of these ranks, and the lower 50% were discarded. This resulted in a set of 500 source sentences to which at least one of the methods had attributed a candidate sentence with relatively high confidence. Agreement between methods after this step was 28.6%, 18.6% and 13.9% respectively (a uniform 50% increase).

English-speaking subjects were invited to participate in an annotation process via a website. Upon visiting the site subjects were presented with a source sentence, and its set of "similar" candidate sentences as chosen by the four approaches, presented in a random order. In cases where methods had agreed, fewer than eight sentences were displayed (i.e. there were no visible repetitions). The annotators were asked to identify the *two* candidate sentences which were "*most similar in meaning"* to the source sentence, and to award an explicit first and second place accordingly. Upon completing a question, participants progressed onto another selected at random from those having received the fewest annotations so far. Participants were required to identify two sentences in every case, no matter how relevant they thought their meanings were in absolute terms, but were free to cease answering questions at any point. No knowledge of the methods used to generate or select the sentences, or of the purpose of the study, was made available to the annotators.

Approximately 90 mostly native English speakers participated in the annotation process. The number of questions answered by each annotator followed a roughly geometric distribution, with maximum, median and minimum of 266, 8 and 1 respectively. The median time taken to answer each question was 24 seconds. Average Kappa for random pairs of responses was 0.25 for annotators' first choices alone, and 0.39 when first and second choices are treated equally. As we are gathering psycholinguistic data, and not developing a gold standard for a supposed underlying objective classification, such moderate levels of agreement are not problematic. What is important for our purposes is that, given the number of annotators involved, the observed levels of agreement are highly significant. The distribution of agreement levels was more-or-less uniform, with a slight dip in the mid-range. Interestingly there was negligible correlation between inter-annotator agreement and the average time taken to answer each question, indicating that seemingly "hard" questions did not take appreciably longer to answer than "easier" questions.

Table 2 presents a summary the agreement between each of the phrasal similarity methods and the votes of the human annotators. Results are separated into annotators' first choices only, and their combined first and second choices. The figures in parentheses are the raw percentage of votes awarded to each method. As there was some corroboration between the methods themselves, these total more than 100% across methods. The figures outside of the parentheses are agreements expressed as a proportion of chance, taking any such corroboration into account. There was only slight variation in the relative balance of scores when stratified according to annotator agreement: the random control unsurprisingly showed an increase at the lowest agreement levels, with BIGRAM and VECTORBAG increasing slightly with agreement, and COMPEXP peaking in the midrange.

	BIGRAM	VECTORBAG	COMPEXP	RANDOM
1st choices only	3.14 (47%)	**3.29** (49%)	2.29 (35%)	0.60 (8%)
All votes	2.87 (43%)	**3.03** (45%)	2.21 (33%)	0.74 (11%)

Table 2: Agreement between computational phrasal similarity measures and human annotations

Despite its ignorance of word order and context, the most successful method in this experiment is the bag of vectors. Arguably more remarkable is the success of the relatively naïve string similarity measure. The fact that these methods also show a surprisingly high degree of agreement with each other (28.6%), suggests that a fair proportion of the phrase similarities present in our dataset can be adequately identified simply by the word forms that comprise them, without recourse to distributional information. Our compositional expectation model is less successful overall, though still receiving several times as many votes as the random control. The fact that its agreement with the two baselines is comparatively low would suggest that it is identifying a different kind of similarity. Given that the mechanisms of compositional expectation are least understood, some kind of qualitative analysis may provide useful insight. To this end, tables 3 and 4 show a selection of 10 phrases deemed *most* similar by the COMPEXP model, that were unanimously awarded first place by the human annotators or unanimously rejected respectively. Rejected phrases are only shown for cases where there was a clear favourite phrase which had been selected by one of the competing models (also shown). The examples were hand-picked for illustrative purposes from qualifying lists of two to three times the size.

The phrase pairs in table 3 can be broadly categorized in terms of their structural and semantic similarities. While one can observe cases of phrases which have near-synonymous meanings in spite

of markedly different wording or structure (B, D, F, G, I, J), there also seem to exist pairs which have essentially equivalent structures, yet are somewhat more loosely related in meaning (A, C, E, H). Note that this is a very informal analysis and a lot of overlap between these classes can be acknowledged.

	Source phrase	**Selected phrase**
A	She moved cautiously into the room	She looked slowly around the room
B	She's an exceptionally nice woman	She's really a nice person
C	It was a desperately lonely time	It was a really bad time
D	Of course I take it seriously	I took it terribly seriously
E	He went into the sitting room	He entered the throne room
F	I left the room	I ran back out of the room
G	He held desperately onto her arm	He held her tightly
H	She hurried towards the white van	She ran straight out of the house
I	I'd hardly made a sound	I could manage only a whisper
J	He made his reasons absolutely clear	He certainly made his point

Table 3: Phrases uniquely selected by compositional expectation model and unanimously selected by annotators.

	Source phrase	**Rejected phrase**	**Most strongly selected phrase**
A	Good, good, good	Sweet and beautiful and good	Good good good
B	It was a really good night	It makes a good story	I thought it was really good
C	He moved slowly along the beach	He moved vaguely around the room	He moved slowly and quietly
D	It was peaceful by the river	It was dark in the room	They even took to the river
E	The event went smoothly and pleasantly	The house was dark and quiet	It was a good time really
F	They took it very badly	Obviously they'd lost it	My family took it badly
G	He really should have won it	He took it personally	It's good we won
H	I take the left	I ran back out of the room	Take a big left turn
I	He's made a good marriage	He's probably making a mistake	He made a good world
J	Isobel moved restlessly around the room	Gaily moved it nearer the counter	She looked slowly around the room

Table 4: Phrases uniquely selected by compositional expectation model, but unanimously rejected by annotators.

Compared to table 3, those phrases in table 4 which exhibit similar structure relate more loosely in their subject matter (C, D, E, J); nonetheless, parts-of-speech and aspects of semantic category do seem to be largely preserved (*dark-peaceful, room-river-beach, in-by, along-around slowly-vaguely, Isobel-Gaily*) resulting in some cases in what might better be described as analogy than synonymy. Where phrases do deviate in structure, their similarities are much less prescriptive; in many cases they seem to imply an almost rhetorical relationship (B, F, G, I).

These observations seem to indicate that compositional expectation is capable of capturing both structural and semantic aspects of similarity, with a leaning towards the former. With some notable exceptions (E), the phrases chosen by the competing methods - and preferred by the annotators - tend to exhibit a more literal or topical relationship with the source phrase (in keeping with how these methods are formulated). We should point out though that while the qualitative observations made here do seem to hold in some measure across the dataset, it is very difficult - and necessarily left to future work - to objectify them; it remains possible that the some of the patterns identified are due to chance and the limited set of phrases comprising our dataset.

5. Discussion

We have presented a novel approach to purely distributional semantic composition, called compositional expectation. By employing a language model and representing phrase constituents in terms of the expectations evoked in their stead, it is possible to represent phrase meaning in a way that is sensitive to word order without recourse to problematic vector operations. While there is evidence to suggest that this approach is able to usefully capture compositional meaning in certain cases – and in a manner that is complementary to more naive methods - its overall performance as measured against the two human-annotated datasets in this study suggests a lot of room for improvement. At present it is unclear to what extent this is a reflection of limitations of these datasets (whether either of them

accurately models the problem), deficiencies in the specific formulation of our model (the language model employed, the vector handling etc), or fundamental shortcomings of this approach to capturing semantic similarity.

A qualitative analysis seems to suggest that compositional expectation is capable of capturing some quite complex structural similarities, as well as broad semantic correspondences between dissimilarly structured phrases. While some capacity to encode structure was presupposed owing to the strong dependence of expectation vectors upon the context in which words appear, it does not obviously follow that similar-meaning but structurally-unlike phrases should have comparable vectors. Our vectors are simply calculated on a token-by-token basis, without consideration of any hierarchical structure present - a generally assumed requirement of compositional models (Padó & Lapata, 2007; Mitchell & Lapata, 2010). While we can speculate that the expectations generated at the terminal positions of synonymous phrases (and therefore sub-phrases) ought to be similar, it is hard to imagine that the strong internal expectations within idiomatic expressions, say, are anything but obstructive to compositional meaning. Perhaps we will find that, as with Schütze's (1998) second-order vectors, the information associated with the most pertinent interpretation tends to dominate the sum. If this is so then the fact that simple vector addition is both associative and commutative - and therefore agnostic of any structure present - may actually play an important role in these models. Given this commutative operator, there would seem to be no means of determining retrospectively to which constituents of a phrase any portion of a compositional vector belongs; this would appear to be a fundamental limitation with respect to the encoding of structure. However, we can speculate that in practice the overwhelming contingent of possible factorizations will tend to be syntactically or semantically implausible. The most plausible interpretations may therefore tend to be those formed by the original word vectors, or very similar ones. Because under an expectation model, jumbling the word order tends to result in *different* component vectors, most ordered interpretations of such factorizations will also tend to be implausible (unless perhaps they actually constitute a valid paraphrase). As this kind of plausibility is information which the language user has, it need not be encoded.

Such lines of thought suggest another problem which needs to be addressed if any of the models considered herein are to be claimed as cognitively plausible takes on compositional meaning: the re-encoding of vector representations into natural language. This would seem to be a straightforward but computationally hard search problem, analogous to that which lies at the heart of machine translation: simultaneously maximizing the plausibility and the faithfulness of a linguistic realisation. Indeed, this might be the acid-test for any proposed compositional representation of meaning. If such "language-meaning codecs" are attainable, then it would pave the way for a host of applications that work natively with meaning, and could revolutionize the way search engines, dialogue agents and machine translation systems are engineered.

Acknowledgements

Sincerest thanks are extended to Jeff Mitchell and Mirella Lapata for the use of their dataset, and as ever to Eric Atwell and Katja Markert for their constructive criticism and support.

Bibliography

Palakorn Achananuparp (2008), "The Evaluation of Sentence Similarity Measures", Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery

F. Attneave (1959), "Applications of Information Theory to Psychology: A summary of basic concepts, methods, and results". Holt, Rinehart and Winston.

Baroni (2010) "Distributional semantics IV: Is distributional semantics really 'semantics'?", UPF Computational Semantics Course

Harm Brouwer, Hartmut Fitz & John C. J. Hoeks (2010), "Modeling the Noun Phrase versus Sentence Coordination Ambiguity in Dutch: Evidence from Surprisal Theory" Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics, ACL 2010, pages 72–80,

Stephen Clark, Bob Coecke & Mehrnoosh Sadrzadeh (2008), "A Compositional Distributional Model of Meaning", Proceedings of the Second Symposium on Quantum Interaction (QI-2008), pp.133-140

Katrin Erk & Sebastian Padó (2008), "A structured vector space model for word meaning in context", Proceedings of EMNLP 2008.

Katrin Erk & Sebastian Padó (2009). "Paraphrase assessment in structured vector space: Exploring parameters and datasets". Proceedings of the EACL Workshop on Geometrical Methods for Natural Language Semantics

Pablo Gamallo, Gabriel P Lopes & Alexandre Agustini (2004) "The Role of Optional Co-composition to Solve Lexical and Syntactic Ambiguity", Procesamiento del Lenguaje Natural, volume 33, pages 73-80

Francoise Gayral, Nathalie Pernelle1 & Patrick Saint-Dizier Gayral (2000), "On Verb Selectional Restrictions: Advantages and Limitations", NLP 2000, LNCS 1835, pages. 57-68

Thomas K Landauer & Susan T Dumais (1997), "A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge." Psychological Review CIV/2. 211-240.

M. Lesk. (1986). "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from and ice cream cone". In Proceedings of the ACM SIGDOC Conference, pages 24–26, Toronto, Canada.

Jones & Mewhort (2007), "Representing word meaning and order information in a composite holographic lexicon." Psychological Review, 114, 3 1–37.

Walter Kintsch & Praful Mangalath (2010), "The Construction of Meaning", Topics in Cognitive Science

Jeff Mitchell, Mirella Lapata, Vera Demberg & Frank Keller (2010), "Syntactic and Semantic Factors in Processing Difficulty: An Integrated Measure" Proceedings of ACL 2010

Jeff Mitchell & Mirella Lapata (2008), "Vector-based models of semantic composition". Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 236-244.

Jeff Mitchell & Mirella Lapata (2010). "Composition in Distributional Models of Semantics". Cognitive Science (to appear).

Sebastian Padó & Mirella Lapata (2007). "Dependency-based construction of semantic space models". Computational Linguistics XXXIII/2. 161-199.

Plate (1995),"Holographic reduced representations". IEEE Transactions on Neural Networks, 6, 623–641.

James Putsejovsky (1995), "The Generative Lexicon." Cambridge, MA: MIT Press.

Sebastian Rudolph & eugenie Giesbrecht (2010), "Compositional Matrix-Space Models of Language", Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 907-916

Magnus Sahlgren, Anders Holst, & Pentti Kanerva (2008), "Permutations as a means to encode order in word space". Proceedings of Cognitive Science 2008, pages 1300–1305.

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

Paul Smolensky (1990). "Tensor product variable binding and the representation of symbolic structures in connectionist networks". Artificial Intelligence, 46, 159–216.

Stefan Thater, Hagen Fürstenau & Manfred Pinkal (2010). "Contextualizing Semantic Represenations Using Syntactically Enriched Vector Models". Proceedings of ACL2010

Justin Washtell (2009). "Co-dispersion: A windowless approach to lexical association." Proceedings EACL'09.

Justin Washtell (2010). "Expectation Vectors: A Semiotics-Inspired Approach to Geometric Lexical-Semantic Representation", GEMS-2010

Justin Washtell & Katja Markert (2009). "Comparing windowless and window-based computational association measures as predictors of syntagmatic human associations". In Proceedings of EMNLP-2009, pages 628-637.

Deniz Yuret (2007), "KU: Word Sense Disambiguation by Substitution", Proceedings of SemEval-2007

Deniz Yuret (2010), "The Noisy Channel Model for Unsupervised Word Sense Disambiguation, Computational Linguistics", Volume 31, Number 1