Sentence paraphrase detection: When determiners and word order make the difference

Nghia Pham
University of Trento
thenghia.pham@unitn.it

Raffaella Bernardi
University of Trento
bernardi@disi.unitn.it

Yao Zhong Zhang
The University of Tokyo
zhangyaoz@gmail.com

Marco Baroni
University of Trento
marco.baroni@unitn.it

Abstract

Researchers working on distributional semantics have recently taken up the challenge of going beyond lexical meaning and tackle the issue of compositionality. Several Compositional Distributional Semantics Models (CDSMs) have been developed and promising results have been obtained in evaluations carried out against data sets of small phrases and as well as data sets of sentences. However, we believe there is the need to further develop good evaluation tasks that show whether CDSM truly capture compositionality. To this end, we present an evaluation task that highlights some differences among the CDSMs currently available by challenging them in detecting semantic differences caused by word order switch and by determiner replacements. We take as starting point simple intransitive and transitive sentences describing similar events, that we consider to be paraphrases of each other but not of the foil paraphrases we generate from them. Only the models sensitive to word order and determiner phrase meaning and their role in the sentence composition will not be captured into the foils' trap.

1 Introduction

Distributional semantics models (DSMs) have recently taken the challenge to move up from lexical to compositional semantics. Through many years of almost exclusive focus on lexical semantics, many data sets have been developed to properly evaluate which aspects of lexical meaning and lexical relations are captured by DSMs. For instance, DSMs have been shown to obtain high performance in simulating semantic priming (Lund and Burgess, 1996), predicting semantic similarity (McDonald, 2000) and association (Griffiths et al., 2007) and have been shown to achieve human level performance on synonymy tests such as those included in the Test of English as a Foreign Language (TOEFL) (Landauer and Dumais, 1997). Compositional DSMs (CDSMs) are of more recent birth, and thus their proponents have focused effort on the study of the compositional operations that are mathematically available and empirically justifiable in vector-space models. Important progress has been made and several models have now been implemented ranging from the additive and multiplicative models of Mitchell and Lapata (2010), to functional models based on tensor contraction (Clark, 2012; Coecke et al., 2010; Baroni and Zamparelli, 2010), to the one based on recursive neural networks of Socher et al. (2011). We believe it is now necessary to shift focus somewhat to the semantic tasks against which to evaluate these models, and to develop appropriate data sets to better understand which aspects of natural language compositionality we are already capturing, what could still be achieved and what might be beyond the scope of this framework. This paper tries to contribute to this new effort. To this end, we start by looking at data sets of phrases and sentences used so far to evaluate CDSMs.

Word order in phrase similarity Starting from a data set of pairs of noun-noun, verb-noun and adjective-noun phrases (e.g., *certain circumstance* and *particular case*) rated by humans with respect

to similarity (Mitchell and Lapata, 2010), Turney (2012) obtains an extended version including word order variations of the original phrases, which are automatically judged to have a very low similarity (e.g., *certain circumstance* and *case particular*).

Sentence similarity: Intransitive Sentences One of the first proposals to look at verb-argument composition traces back to Kintsch (2001) who was interested in capturing the different verb senses activated by different arguments, e.g., "*The color ran*" vs. "*The horse ran*", but the model was tested only on a few sentences. Starting from this work, Mitchell and Lapata (2008) made an important step forward by developing a larger data set of subject+intransitive-verb sentences. They began with frequent noun-verb tuples (e.g., *horse ran*) extracted from the British National Corpus (BNC) and paired them with sentences with two synonyms of the verb, representing distinct verb senses, one compatible and the other incompatible with the argument (e.g., *horse galloped* and *horse dissolved*). The tuples were converted into simple sentences (in past tense form) and articles were added to nouns when appropriate. The final data set consists of 120 sentences with 15 original verbs each composed with two subject nouns and paired with two synonyms. Sentence pair similarity was rated by 49 volunteers on the web.

Sentence similarity: Transitive Sentences Following the method proposed in Mitchell and Lapata (2008), Grefenstette and Sadrzadeh (2011b) developed an analogous data set of transitive sentences. Again the focus is on how arguments (subjects and objects) influence the selection of the meaning of an ambiguous verb. For instance, *meet* is synonymous both of *satisfy* and *visit*. For each verb (in total 10 verbs), 10 subject+transitive-verb+object tuples with the given verb were extracted from the BNC and sentences in simple past form (with articles if necessary) were generated. For example, starting from *met*, the two sentences "*The system met the criterion*" and "*The child met the house*" were generated. For each sentence, two new versions were created by replacing the verb with two synonyms representing two verb senses (e.g., "*The system visited the criterion*" and "*The system satisfied the criterion*"). The data set consists of 200 pairs of sentences annotated with human similarity judgments.

Large-scale full sentence paraphrasing data Socher et al. (2011) and Blacoe and Lapata (2012) tackle the challenging task of evaluating CDSMs against large-scale full sentence data. They use the Microsoft Research Paraphrase Corpus (Dolan et al., 2004) as data set. The corpus consists of 5800 pairs of sentences extracted from news sources on the web, along with human annotations indicating whether each pair captures a paraphrase/semantic equivalence relationship.

The evaluation experiments conducted against these data sets seem to support the following conclusions:

- "the model should be sensitive to the order of the words in a phrase (for composition) or a word pair (for relations), when the order affects the meaning." (Turney, 2012)
- "experimental results demonstrate that the multiplicative models are superior to the additive alternatives when compared against human judgments [about sentence similarity]." (Mitchell and Lapata, 2008)
- "shallow approaches are as good as more computationally intensive alternatives [in sentence paraphrase detection]. They achieve considerable semantic expressivity without any learning, sophisticated linguistic processing, or access to very large corpora." (Blacoe and Lapata, 2012)

With this paper, we want to put these conclusions on stand-by by asking the question of whether the appropriate tasks have really been tackled. The first conclusion above regarding word order is largely shared, but still no evaluation of CDSMs against sentence similarity considers word order seriously. We do not exclude that in real-world tasks systems which ignore word order may still attain satisfactory results (as the results of Blacoe and Lapata 2012 suggest), but this will not be evidence of having truly captured compositionality.

Moreover, a hidden conclusion (or, better, assumption!) of the evaluations conducted so far on CDSMs seems to be that grammatical words, in particular determiners, play no role in sentence meaning and hence sentence similarity and paraphrase detection. A first study on this class of words has been presented in Baroni et al. (2012) where it is shown that DSMs can indeed capture determiner meaning and their role in the entailment between quantifier phrases. The data sets used in Mitchell and Lapata (2008) and Grefenstette and Sadrzadeh (2011b) focus on verb meaning and its sense disambiguation within context, and consider sentences where determiners are just place-holders to simply guarantee grammaticality, but do not play any role neither in the human judgments nor in the model evaluation – in which they are simply ignored. Similarly, Blacoe and Lapata (2012) evaluate the compositional models on full sentences but again ignore the role of grammatical words that are treated as "stop-words". Should we conclude that from a distributional semantic view "*The system met the criterion*", "*No system met the criterion*" and "*Neither system met the criterion*" boil down to the same meaning? This is a conclusion we cannot exclude but neither accept *a-priori*. To start considering these questions more seriously, we built a data set of intransitive and transitive sentences in which word order and determiners have the chance to prove their worth in sentence similarity and paraphrase detection.

2 Compositional Distributional Semantic Models

In this section we won't present a proper overview of CDSMs, but focus only on those models we will be testing in our experiments, namely the multiplicative and additive models of Mitchell and Lapata (2008, 2009, 2010), and the lexical function model that represents the work carried out by Baroni and Zamparelli (2010), Grefenstette and Sadrzadeh (2011b), Grefenstette et al. (2013). We leave a re-implementation of Socher et al. (2012), another approach holding much promise for distributional composition with grammatical words, to future work.

Multiplicative and additive models While Mitchell and Lapata (2008, 2009, 2010) propose a general framework that encompasses most of the CDSMs currently available, their empirical work focuses on two simple but effective models where the components of the vector resulting from the composition of two input vectors contain (functions of) geometric or additive averages of the corresponding input components.

Given input vectors $\mathbf{u}$ and $\mathbf{v}$, the multiplicative model (`mult`) returns a composed vector $\mathbf{p}$ such that each of its components p_i is given by the product of the corresponding input components:

$$p_i = u_i v_i$$

In the additive model (`add`), the composed vector is a sum of the two input vectors:[1]

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

Mitchell and Lapata do not address composition with grammatical words directly, but their approach is obviously aimed at capturing composition between content words.

Lexical function model Baroni and Zamparelli (2010) take inspiration from formal semantics to characterize composition in terms of *function application*. They model adjective-noun phrases by treating the adjective as a function from nouns onto (modified) nouns. Given that linear functions can be expressed by matrices and their application by matrix-by-vector multiplication, a functor (such as the adjective) is represented by a matrix $\mathbf{U}$ to be composed with the argument vector $\mathbf{v}$ (e.g., the noun vector) by multiplication, to return the vector representing the phrase:

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

[1] Mitchell and Lapata also propose two weighted additive models, but it is not clear how to extend them to composition of more than two words.

Non-terminals (Grammar)	Terminals (Lexicon)
$S \rightarrow DP\ VP$	$DET \rightarrow$ a; some; the; one; two; three; no
$DP \rightarrow DET\ N$	$N \rightarrow$ man; lady; violin; guitar; ...
$DP \rightarrow N$	$ADJ \rightarrow$ big; large; acoustic; ...
$N \rightarrow ADJ\ N$	$IV \rightarrow$ performs; drinks; flies; ...
$VP \rightarrow IV$	$TV \rightarrow$ cuts; eats; plays; ...
$VP \rightarrow TV\ DP$	

Figure 1: CFG of the fragment of English in our data set

Adjective matrices are estimated from corpus-extracted examples of input noun vectors and the corresponding output adjective-noun phrase vectors, an idea also adopted by Guevara (2010).

The approach of Baroni and Zamparelli, termed `lexfunc` (because specific *lex*ical items act as *func*tors), is actually a specific instantiation of the DisCoCat formalism (Clark, 2012; Coecke et al., 2010), that looks at the general case of n-ary composition functions, encoded in higher-order tensors, with function application modeled by tensor contraction, a generalization of matrix-by-vector multiplication to tensors of arbitrary order. The DisCoCat approach has also been applied to transitive verbs by Grefenstette and Sadrzadeh (2011a) and Grefenstette and Sadrzadeh (2011b). The regression method proposed in Baroni and Zamparelli (2010) for estimating adjectives has been generalized by Grefenstette et al. (2013) and tested on transitive verbs modeled as two-argument functions (corresponding to third-order tensors).

3 Data set

We have built a data set of transitive and (a few) intransitive sentences that fall within the language recognized by the Context Free Grammar in Figure 1. As the rewriting rules of the grammar show, the subject and (in the case of transitive sentences) the object are always either a determiner phrase built by a determiner and a noun, where the noun can be optionally modified by one or more adjectives, or a bare noun phrase. The verb is always in the present tense and never negated. We use a total of 32 verbs, 7 determiners, 65 nouns, and 19 adjectives.

The data set is split into *paraphrase* and *foil sets*, described below. We use the term "paraphrase" to indicate that two sentences can describe essentially the same situation. The two subsets will be used to test DSMs in a paraphrase detection task, to understand which model better captures compositional meaning in natural language; the foil set, focusing on disruptive word order and determiner changes, has the purpose to help spotting whether a DSM is "cheating" in accomplishing this task, or, better, if it does detect paraphrases but does not properly capture compositionality.

Paraphrase set The sentences are grouped into sets of paraphrases. Some groups are rather similar to each other (for instance they are about someone playing some instrument) though they clearly describe a different situation (the player is a man vs. a woman or the instrument is a guitar vs. a violin), as it happens for the sentences in Group 1 and Group 2 listed in Table 1. We took as starting point the Microsoft Research Video Description Corpus (Chen and Dolan, 2011) considering only those sentences that could be simplified to fit in the CFG described above. We have obtained 20 groups of sentences. Groups which were left with just a few sentences after grammar-based trimming have been extended adding sentences with the nouns modified by an attributive adjective (chosen so that it would not distort the meaning of the sentence, e.g. we have added *tall* as modifier of *person* in "*A tall person makes a cake*", if there is no original sentence in the group that would describe the person differently), or adding sentences with a determiner similar to the one used in the original description (for instance *the* when there was *a*). In total, the set contains 157 sentences, divided into 20 groups; the smallest paraphrase group contains 4 sentences whereas the largest one consists of 17; the groups contain 7.85 sentences on average.

Group 1: Paraphrases	**Group 2: Paraphrases**
P A man plays a guitar	P A girl plays violin
P A man plays an acoustic guitar	P A lady plays violin
P A man plays an electric guitar	P A woman plays violin
P A old man plays guitar	P A woman plays the violin
P The man plays the guitar	
P The man plays music	
P A man plays an instrument	
Group 1: Foils	**Group 2: Foils**
S A guitar plays a man	S A violin plays a girl
S An acoustic guitar plays a man	S A violin plays a lady
S An instrument plays a man	...
D No man plays no guitar	D No girl plays violin
SD No guitar plays no man	D No lady plays violin
SD No guitar plays a man	
D The man plays no guitar	
...	

Table 1: Sample of paraphrases and foils of two groups

Foil set From each group in the paraphrase set, we have obtained foil paraphrase sentences in three ways: (i) by inverting the words in the subject and object position of the original sentences (sentences marked by S in Table 1); (ii) by replacing the determiner with a new one that clearly modifies the meaning of the sentences – replacing the original (positive) determiner with *no* (sentences marked by D); and by inverting subject and object of the sentences obtained by (ii) (sentences marked by SD). Note that the change in determiner, unlike in the case of the true paraphrase set above, is disruptive of meaning compatibility. In total there are 325 foils, the smallest group has 4 foils whereas the largest one consists of 36; on average the foil groups contain 17 sentences each.

4 Experiments

4.1 Semantic space and composition method implementation

We collect co-occurrence statistics from the concatenation of the ukWaC corpus,[2] a mid-2009 dump of the English Wikipedia[3] and the British National Corpus,[4] for a total of about 2.8 billion tokens. We extracted distributional vectors for the 20K most frequent inflected nouns in the corpus, and all the verbs, adjectives and determiners in the vocabulary of our data set (lemma forms). We adopt a bag-of-word approach, counting co-occurrence with all context words in the same sentence with a target item. Context words consist of the 10K most frequent lemmatized content words (nouns, adjectives, verbs, and adverbs). Raw frequencies are converted into Local Mutual Information scores (Evert, 2005), and the dimensions reduced to 200 by means of the Singular Value Decomposition.

For the *lexfunc* model, we used regression learning based on input and output examples automatically extracted from the corpus, along the lines of Baroni and Zamparelli (2010) and Grefenstette et al. (2013), in order to obtain tensors representing functor words in our vocabulary. Determiners, intransitive verbs and adjectives are treated as one-argument functions (second order tensors, or matrices) from nouns to determiner phrases, from determiner phrases to sentences, and from nouns to nouns, respectively. Transitive verbs are treated as two-argument functions (third order tensors) from determiner phrases to determiner phrases to sentences. For the multiplicative and additive models we consider two versions, one in

[2] http://wacky.sslmit.unibo.it/

[3] http://en.wikipedia.org

[4] http://www.natcorp.ox.ac.uk/

which determiners are ignored (as in Blacoe and Lapata, 2012) and one in which they are not. For *lexfunc* and the additive model, we also look at how normalizing the vectors to unit length before composition (both in training and testing) affects performance (*mult* is not affected by scalar transformations).

4.2 Evaluation methods

We have carried out two experiments. The first is a classic paraphrase detection task, in which CDSMs have to automatically cluster the sentences from the paraphrase set into the ground-truth groups. The second one aims to highlight when the possible good performance in the paraphrase detection task does correspond to true modelling of compositionality, that should be sensitive to word order and disruptive changes in the determiner.

Clustering We have carried out this experiment against the paraphrase set. We used the standard globally-optimized repeated bisecting method as implemented in the widely used CLUTO toolkit (Karypis, 2003), using cosines as distance functions, and accepting all of CLUTO's default values. Performance is measured by *purity*, one of the standard clustering quality measures returned by CLUTO (Zhao and Karypis, 2003). If n_r^i is the number of items from the *i*-th true (gold standard) group that were assigned to the *r*-th cluster, n is the total number of items and k the number of clusters, then: $Purity = \frac{1}{n}\sum_{r=1}^{k}\max_i(n_r^i)$. In the case of perfect clusters, purity will be of 100%; as cluster quality deteriorates, purity approaches 0.

Sentence similarity to paraphrases vs. foils The idea of the second experiment is to measure the similarity of each sentence in the paraphrase set to all other sentences in the same group (i.e., valid paraphrases, P), as well as to sentences in the corresponding foil set (FP). For each sentence in a P group, we computed the mean of the cosine of the sentence with all the sentences in the same ground-truth group (with all the P sentences) (`cos.para`) and the mean of the cosine with all the foil paraphrases (with all the FP sentences, viz. those marked by S, D, SD) of the same group (`cos.foil`). Then, we computed the difference between `cos.para` and `cos.foil` (`diff.para.foil`= `cos.para` - `cos.foil`). Finally, we computed the mean of `diff.para.foil` for all the sentences in the data set. Models which achieve higher values are those that are not captured by the foils' trap, since they are able to distinguish paraphrases from their foils: Only a model that realizes that *A man plays an instrument* is a better paraphrase of *A man plays guitar* than either *A guitar plays a man* or *The man plays no guitar* can be said to truly catch compositional meaning, beyond simple word meaning overlap. To focus more specifically on word order, we will report the same analysis also when considering only the scrambled sentences as foils: `diff.para.scrambled`=`cos.para` - `cos.scrambled`, where the latter are means of the cosine of the paraphrase with all the scrambled sentences (with all the sentences marked by S) of the same group (with no manipulation of the determiners).

4.3 Results

In Table 2 we report the performance of all the models evaluated with the two methods discussed above. Concerning the paraphrase clustering task, we first notice that all models are doing much better than the random baseline, and most of them are also above a challenging word-overlap baseline (challenging because sentences in the same groups do tend to share many words) (Kirschner et al., 2009). By far the highest purity value (0.84) was obtained by normalized *add* without determiners. This confirms that "shallow" approaches are indeed very good for paraphrase detection (Blacoe and Lapata, 2012). Interestingly, the additive model performs quite badly (0.32) if it is not normalized and determiners are not stripped off: A reasonable interpretation is that the determiner vectors tend to be both very long (determiners are very frequent) and uninformative (the same determiners occur in most sentences), so their impact must be dampened. Keeping determiners is also detrimental for the multiplicative model, that in general in our experiment does not perform as well as the additive one. The *lexfunc* model

	Experiment 1	Experiment 2	
Model	Purity	diff.para.foil	diff.para.scrambled
mult	0.49	0.05 (0.21)	0.04 (0.29)
mult no det	0.62	0.00 (0.19)	-0.01 (0.23)
add	0.32	0.12 (0.09)	0.00 (0.07)
add norm	0.78	0.06 (0.05)	0.00 (0.04)
add no det	0.74	0.00 (0.12)	0.01 (0.11)
add norm no det	**0.84**	0.00 (0.06)	0.00 (0.06)
lexfunc	0.59	**0.24** (0.25)	**0.28** (0.35)
lexfunc norm	0.75	0.06 (0.08)	0.09 (0.11)
word overlap	0.59		
random	0.11		

Table 2: Experiment results (mean diff.para.foil and diff.para.scrambled values followed by standard deviations)

without normalization performs at the level of the word-overlap baseline and the best multiplicative model, whereas its performance after normalization reaches that of *add* without determiners. Note that for *lexfunc*, normalization cannot be a way to lower the impact of determiners (that in this model are matrices, not vectors), so future work must ascertain why we observe this effect.

Coming now to the second experiment, we note that most models fell into the foils' trap. For neither multiplicative models the difference between similarity to true paraphrases vs. foils is significantly above zero (here and below, statistical significance measured by two-tailed t-tests). Among additive models, only those that do include information about determiners have differences between paraphrase and foil similarity significantly above zero. Indeed, since the additive model is by construction insensitive to word order, the fact that it displays a significant difference at all indicates that evidently the vectors representing determiners are more informative about their meaning than we thought. Still, we should remember that the only determiner replacement tested is the one from the positive determiners – *a*, *some*, *the*, *one*, *two*, *three* – to *no*. Further studies on the role of determiners in the distributional meaning of sentences should be carried out, before any strong conclusion can be drawn. Finally, both *lexfunc* models display paraphrase-foil differences significantly above zero, and the non-normalized model in particular works very well in this setting (being also significantly better than the second best, namely the *add* model).

The comparison of the values obtained for diff.para.foil and diff.para.scrambled is only interesting for the *lexfunc* models. Remember that the latter comparison tells us which models fail to compose meaning because they are insensitive to word order, whereas the former also takes determiners into account. Since *mult* and *add* do not take word order into account, they of course have values of diff.para.scrambled that are not significantly different from 0 (consider this a sanity check!). Both *lexfunc* variants have values for this variable that are significantly higher than 0, showing that this model is not only taking word order into account, but making good use of this information.

To conclude, all compositional models perform paraphrase clustering quite well, and indeed on this task a simple additive model performs best. However, the picture changes if instead of considering groups of paraphrases extracted from a standard paraphrase data set, we look at a task where paraphrases must be distinguished by foils that are deliberately constructed to take the effect of determiners and word order into account. In this setup, only the *lexfunc* model is consistently performing above chance level. Still, since the version of *lexfunc* that handles the second task best only performs paraphrase clustering at the level of the word-overlap baseline, we cannot really claim that this is a satisfactory model of compositional sentence meaning, and clearly further work is called for to test and develop better compositional distributional semantic models.

5 Conclusion

We have proposed a new method for evaluating Compositional Distributional Models (CDSMs) that we believe can get a more reliable fingerprint of how different CDSMs are capturing compositionality. Turney (2012) has proposed to create foil paraphrases of phrases by switching the word order (often resulting in ungrammatical sequences). We have extended the method to sentential paraphrases (while guaranteeing grammaticality) and have added a new trap for the CDSMs, namely determiner replacement. Starting from the Microsoft Research video description corpus, we have developed a data set organized in groups of paraphrases and foils (obtained by both word order switch and determiner replacement) and evaluated the performance of several CDSMs against it. None of the models can be claimed to be the successful one, since the additive model is best in capturing paraphrase clustering, whereas the *lexfunc* model is best in distinguishing sentences involving word order switch and the effect of determiners. For the future, we might consider the possibility of investigating an hybrid system that combines insights from these two models. Furthermore, the current data set does not provide enough variety in meaning-preserving and disruptive determiner changes to single out the effect of determiners like we did for word order, hence further studies in this direction are required. All in all, we believe the evaluation confirms the need of setting up semantic tasks suitable for evaluating the real challenges CDSMs are said to be tackling. We hope that the data set we developed can be a step in this direction. To this extent, we make it publicly available from `http://clic.cimec.unitn.it/composes/`.

Acknowledgments

The work has been funded by the ERC 2011 Starting Independent Research Grant supporting the COMPOSES project (nr. 283554).

References

Baroni, M., R. Bernardi, N.-Q. Do, and C.-C. Shan (2012). Entailment above the word level in distributional semantics. In *Proceedings of EACL*, Avignon, France, pp. 23–32.

Baroni, M. and R. Zamparelli (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, Boston, MA, pp. 1183–1193.

Blacoe, W. and M. Lapata (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of EMNLP*, Jeju Island, Korea, pp. 546–556.

Chen, D. L. and W. B. Dolan (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of ACL*, Portland, OR, pp. 190–200.

Clark, S. (2012). Type-driven syntax and semantics for composing meaning vectors. In C. Heunen, M. Sadrzadeh, and E. Grefenstette (Eds.), *Quantum Physics and Linguistics: A Compositional, Diagrammatic Discourse*. Oxford, UK: Oxford University Press. In press.

Coecke, B., M. Sadrzadeh, and S. Clark (2010). Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis 36*, 345–384.

Dolan, W., C. Quirk, and C. Brockett (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of COLING*, pp. 350–356.

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

Grefenstette, E., G. Dinu, Y.-Z. Zhang, M. Sadrzadeh, and M. Baroni (2013). Multi-step regression learning for compositional distributional semantics. In *Proceedings of IWCS*, Potsdam, Germany. In press.

Grefenstette, E. and M. Sadrzadeh (2011a). Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of EMNLP*, Edinburgh, UK, pp. 1394–1404.

Grefenstette, E. and M. Sadrzadeh (2011b). Experimenting with transitive verbs in a DisCoCat. In *Proceedings of GEMS*, Edinburgh, UK, pp. 62–66.

Griffiths, T., M. Steyvers, and J. Tenenbaum (2007). Topics in semantic representation. *Psychological Review 114*, 211–244.

Guevara, E. (2010). A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of GEMS*, Uppsala, Sweden, pp. 33–37.

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

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

Kirschner, M., R. Bernardi, . Baroni, and L. T. Dinh (2009). Analyzing interactive QA dialogues using logistic regression models. In *Proceedings of AI*IA*, pp. 334–344.

Landauer, T. and S. Dumais (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211–240.

Lund, K. and C. Burgess (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods 28*, 203–208.

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

Mitchell, J. and M. Lapata (2008). Vector-based models of semantic composition. In *Proceedings of ACL*, Columbus, OH, pp. 236–244.

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

Mitchell, J. and M. Lapata (2010). Composition in distributional models of semantics. *Cognitive Science 34*(8), 1388–1429.

Socher, R., E. Huang, J. Pennin, A. Ng, and C. Manning (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of NIPS*, Granada, Spain, pp. 801–809.

Socher, R., B. Huval, C. Manning, and A. Ng (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP*, Jeju Island, Korea, pp. 1201–1211.

Turney, P. (2012). Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research 44*, 533–585.

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