A relatedness benchmark to test the role of determiners in compositional distributional semantics

Raffaella Bernardi and **Georgiana Dinu** and **Marco Marelli** and **Marco Baroni**
Center for Mind/Brain Sciences (University of Trento, Italy)
first.last@unitn.it

Abstract

Distributional models of semantics capture word meaning very effectively, and they have been recently extended to account for compositionally-obtained representations of phrases made of content words. We explore whether compositional distributional semantic models can also handle a construction in which grammatical terms play a crucial role, namely determiner phrases (DPs). We introduce a new publicly available dataset to test distributional representations of DPs, and we evaluate state-of-the-art models on this set.

1 Introduction

Distributional semantics models (DSMs) approximate meaning with vectors that record the distributional occurrence patterns of words in corpora. DSMs have been effectively applied to increasingly more sophisticated semantic tasks in linguistics, artificial intelligence and cognitive science, and they have been recently extended to capture the meaning of phrases and sentences via compositional mechanisms. However, scaling up to larger constituents poses the issue of how to handle *grammatical* words, such as determiners, prepositions, or auxiliaries, that lack rich conceptual content, and operate instead as the logical "glue" holding sentences together.

In typical DSMs, grammatical words are treated as "stop words" to be discarded, or at best used as context features in the representation of *content* words. Similarly, current *compositional* DSMs (cDSMs) focus almost entirely on phrases made of two or more content words (e.g., adjective-noun or verb-noun combinations) and completely ignore grammatical words, to the point that even the test set of transitive sentences proposed by Grefenstette and Sadrzadeh (2011) contains only Tarzan-style statements with determiner-less subjects and objects: *"table show result"*, *"priest say mass"*, etc. As these examples suggest, however, as soon as we set our sight on modeling phrases and sentences, grammatical words are hard to avoid. Stripping off grammatical words has more serious consequences than making you sound like the Lord of the Jungle. Even if we accept the view of, e.g., Garrette et al. (2013), that the logical framework of language should be left to other devices than distributional semantics, and the latter should be limited to similarity scoring, still ignoring grammatical elements is going to dramatically distort the very similarity scores (c)DSMs should provide. If we want to use a cDSM for the classic similarity-based paraphrasing task, the model shouldn't conclude that *"The table shows **many** results"* is identical to *"the table shows **no** results"* since the two sentences contain the same content words, or that *"to kill **many** rats"* and *"to kill **few** rats"* are equally good paraphrases of *"to exterminate rats"*.

We focus here on how cDSMs handle *determiners* and the phrases they form with nouns (*determiner phrases*, or DPs).[1] While determiners are only a subset of grammatical words, they are a large and important subset, constituting the natural stepping stone towards sentential distributional semantics: Compositional methods have already been successfully applied to simple noun-verb and noun-verb-noun structures (Mitchell and Lapata, 2008; Grefenstette and Sadrzadeh, 2011), and determiners are just what is missing to turn these skeletal constructions into full-fledged sentences. Moreover, determiner-noun phrases are, in superficial syntactic terms, similar to the adjective-noun phrases that have already been extensively studied from a cDSM perspective by Baroni and Zampar-

[1] Some linguists refer to what we call DPs as noun phrases or NPs. We say DPs simply to emphasize our focus on determiners.

elli (2010), Guevara (2010) and Mitchell and Lapata (2010). Thus, we can straightforwardly extend the methods already proposed for adjective-noun phrases to DPs.

We introduce a new task, a similarity-based challenge, where we consider nouns that are strongly conceptually related to certain DPs and test whether cDSMs can pick the most appropriate related DP (e.g., *monarchy* is more related to *one ruler* than *many rulers*).[2] We make our new dataset publicly available, and we hope that it will stimulate further work on the distributional semantics of grammatical elements.[3]

2 Composition models

Interest in compositional DSMs has skyrocketed in the last few years, particularly since the influential work of Mitchell and Lapata (2008; 2009; 2010), who proposed three simple but effective composition models. In these models, the composed vectors are obtained through component-wise operations on the constituent vectors. Given input vectors $\mathbf{u}$ and $\mathbf{v}$, the multiplicative model (**mult**) returns a composed vector $\mathbf{p}$ with: $p_i = u_i v_i$. In the weighted additive model (**wadd**), the composed vector is a weighted sum of the two input vectors: $\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v}$, where α and β are two scalars. Finally, in the **dilation** model, the output vector is obtained by first decomposing one of the input vectors, say $\mathbf{v}$, into a vector parallel to $\mathbf{u}$ and an orthogonal vector. Following this, the parallel vector is dilated by a factor λ before re-combining. This results in: $\mathbf{p} = (\lambda - 1)\langle \mathbf{u}, \mathbf{v}\rangle \mathbf{u} + \langle \mathbf{u}, \mathbf{u}\rangle \mathbf{v}$.

A more general form of the additive model (**fulladd**) has been proposed by Guevara (2010) (see also Zanzotto et al. (2010)). In this approach, the two vectors to be added are pre-multiplied by weight matrices estimated from corpus-extracted examples: $\mathbf{p} = \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v}$.

Baroni and Zamparelli (2010) and Coecke et al. (2010) take inspiration from formal semantics to characterize composition in terms of *function application*. The former 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 multiplied with the argument vector $\mathbf{v}$ (e.g., the noun vector): $\mathbf{p} = \mathbf{U}\mathbf{v}$. Adjective matrices are estimated from corpus-extracted examples of noun vectors and corresponding output adjective-noun phrase vectors, similarly to Guevara's approach.[4]

3 The noun-DP relatedness benchmark

Paraphrasing a single word with a phrase is a natural task for models of compositionality (Turney, 2012; Zanzotto et al., 2010) and determiners sometimes play a crucial role in defining the meaning of a noun. For example a *trilogy* is composed of ***three*** *works*, an *assemblage* includes ***several*** *things* and an *orchestra* is made of ***many*** *musicians*. These examples are particularly interesting, since they point to a "conceptual" use of determiners, as components of the stable and generic meaning of a content word (as opposed to situation-dependent deictic and anaphoric usages): for these determiners the boundary between content and grammatical word is somewhat blurred, and they thus provide a good entry point for testing DSM representations of DPs on a classic similarity task. In other words, we can set up an experiment in which having an effective representation of the determiner is crucial in order to obtain the correct result.

Using regular expressions over WordNet glosses (Fellbaum, 1998) and complementing them with definitions from various online dictionaries, we constructed a list of more than 200 nouns that are strongly conceptually related to a specific DP. We created a multiple-choice test set by matching each noun with its associated DP (*target DP*), two "foil" DPs sharing the same noun as the target but combined with other determiners (*same-N foils*), one DP made of the target determiner combined with a random noun (*same-D foil*), the target determiner (*D foil*), and the target noun (*N foil*). A few examples are shown in Table 1. After the materials were checked by all authors, two native speakers took the multiple-choice test. We removed the cases (32) where these subjects provided an unexpected answer. The final set,

[2] Baroni et al. (2012), like us, study determiner phrases with distributional methods, but they do not model them compositionally.

[3] Dataset and code available from `clic.cimec.unitn.it/composes`.

[4] Other approaches to composition in DSMs have been recently proposed by Socher et al. (2012) and Turney (2012). We leave their empirical evaluation on DPs to further work, in the first case because it is not trivial to adapt their complex architecture to our setting; in the other because it is not clear how Turney would extend his approach to represent DPs.

noun	*target DP*	*same-N foil 1*	*same-N foil 2*	*same-D foil*	*D foil*	*N foil*
duel	two opponents	various opponents	three opponents	two engineers	two	opponents
homeless	no home	too few homes	one home	no incision	no	home
polygamy	several wives	most wives	fewer wives	several negotiators	several	wives
opulence	too many goods	some goods	no goods	too many abductions	too many	goods

Table 1: Examples from the noun-DP relatedness benchmark

characterized by full subject agreement, contains 173 nouns, each matched with 6 possible answers. The target DPs contain 23 distinct determiners.

4 Setup

Our semantic space provides distributional representations of determiners, nouns and DPs. We considered a set of 50 determiners that include all those in our benchmark and range from quantifying determiners (*every*, *some*...) and low numerals (*one* to *four*), to multi-word units analyzed as single determiners in the literature, such as *a few*, *all that*, *too much*. We picked the 20K most frequent nouns in our source corpus considering singular and plural forms as separate words, since number clearly plays an important role in DP semantics. Finally, for each of the target determiners we added to the space the 2K most frequent DPs containing that determiner and a target noun.

Co-occurrence statistics were collected from the concatenation of ukWaC, a mid-2009 dump of the English Wikipedia and the British National Corpus,[5] with a total of 2.8 billion tokens. We use a bag-of-words approach, counting co-occurrence with all context words in the same sentence with a target item. We tuned a number of parameters on the independent MEN word-relatedness benchmark (Bruni et al., 2012). This led us to pick the top 20K most frequent content word lemmas as context items, Pointwise Mutual Information as weighting scheme, and dimensionality reduction by Non-negative Matrix Factorization.

Except for the parameter-free *mult* method, parameters of the composition methods are estimated by minimizing the average Euclidean distance between the model-generated and corpus-extracted vectors of the 20K DPs we consider.[6] For the *lexfunc* model, we assume that the determiner is the functor and the noun is the argument, and estimate separate matrices representing each determiner using the 2K DPs in the semantic space that contain that determiner. For *dilation*, we treat direction of stretching as a parameter, finding that it is better to stretch the noun.

Similarly to the classic TOEFL synonym detection challenge (Landauer and Dumais, 1997), our models tackle the relatedness task by measuring cosines between each target noun and the candidate answers and returning the item with the highest cosine.

method	*accuracy*	*method*	*accuracy*
lexfunc	39.3	noun	17.3
fulladd	34.7	random	16.7
observed	34.1	mult	12.7
dilation	31.8	determiner	4.6
wadd	23.1		

Table 2: Percentage accuracy of composition methods on the relatedness benchmark

5 Results

Table 2 reports the accuracy results (mean ranks of correct answers confirm the same trend). All models except *mult* and *determiner* outperform the trivial *random* guessing baseline, although they are all well below the 100% accuracy of the humans who took our test. For the *mult* method we observe a very strong bias for choosing a single word as answer (>60% of the times), which in the test set is always incorrect. This leads to its accuracy being below the chance level. We suspect that the highly "intersective" nature of this model (we obtain very sparse composed DP vectors, only ≈4% dense) leads to it not being a reliable method for comparing sequences of words of different length: Shorter sequences will be considered more similar due to their higher density. The *determiner*-only baseline (using the vector of the component determiner as surrogate for the DP) fails because D vectors tend to be far from N vectors, thus the N foil is often preferred to the correct response (that is represented, for this baseline, by its D). In the *noun*-only baseline (use the vector of the component noun as surrogate for the DP),

[5] `wacky.sslmit.unibo.it`; `www.natcorp.ox.ac.uk`

[6] All vectors are normalized to unit length before composition. Note that the objective function used in estimation minimizes the distance between model-generated and corpus-extracted vectors. We do *not* use labeled evaluation data to optimize the model parameters.

the correct response is identical to the same-N and N foils, thus forcing a random choice between these. Not surprisingly, this approach performs quite badly. The *observed* DP vectors extracted directly from the corpus compete with the top compositional methods, but do not surpass them.[7]

The *lexfunc* method is the best compositional model, indicating that its added flexibility in modeling composition pays off empirically. The *fulladd* model is not as good, but also performs well. The *wadd* and especially *dilation* models perform relatively well, but they are penalized by the fact that they assign more weight to the noun vectors, making the right answer dangerously similar to the same-N and N foils.

Taking a closer look at the performance of the best model (*lexfunc*), we observe that it is not equally distributed across determiners. Focusing on those determiners appearing in at least 4 correct answers, they range from those where *lexfunc* performance was very significantly above chance ($p<0.001$ of equal or higher chance performance): *too few*, *all*, *four*, *too much*, *less*, *several*; to those on which performance was still significant but less impressively so ($0.001 < p < 0.05$): *several*, *no*, *various*, *most*, *two*, *too many*, *many*, *one*; to those where performance was not significantly better than chance at the 0.05 level: *much*, *more*, *three*, *another*. Given that, on the one hand, performance is not constant across determiners, and on the other no obvious groupings can account for their performance difference (compare the excellent *lexfunc* performance on *four* to the lousy one on *three*!), future research should explore the contextual properties of specific determiners that make them more or less amenable to be captured by compositional DSMs.

6 Conclusion

DSMs, even when applied to phrases, are typically seen as models of content word meaning. However, to scale up compositionally beyond the simplest constructions, cDSMs must deal with grammatical terms such as determiners. This paper started exploring this issue by introducing a new and publicly available set testing DP semantics in a similarity-based task and using it to systematically evaluate, for the first time, cDSMs on a construction involving grammatical words. The most important take-home message is that distributional representations are rich enough to encode information about determiners, achieving performance well above chance on the new benchmark.

Theoretical considerations would lead one to expect a "functional" approach to determiner representations along the lines of Baroni and Zamparelli (2010) and Coecke et al. (2010) to outperform those approaches that combine vectors separately representing determiners and nouns. This prediction was largely borne out in the results, although the additive models, and particularly *fulladd*, were competitive rivals.

We attempted to capture the distributional semantics of DPs using a fairly standard, "vanilla" semantic space characterized by latent dimensions that summarize patterns of co-occurrence with content word contexts. By inspecting the context words that are most associated with the various latent dimensions we obtained through Non-negative Matrix Factorization, we notice how they are capturing broad, "topical" aspects of meaning (the first dimension is represented by *scripture*, *believer*, *resurrection*, the fourth by *fever*, *infection*, *infected*, and so on). Considering the sort of semantic space we used (which we took to be a reasonable starting point because of its effectiveness in a standard lexical task), it is actually surprising that we obtained the significant results we obtained. Thus, a top priority in future work is to explore different contextual features, such as adverbs and grammatical terms, that might carry information that is more directly relevant to the semantics of determiners.

Another important line of research pertains to improving composition methods: Although the best model, at 40% accuracy, is well above chance, we are still far from the 100% performance of humans. We will try, in particular, to include non-linear transformations in the spirit of Socher et al. (2012), and look for better ways to automatically select training data.

Last but not least, in the near future we would like to test if cDSMs, besides dealing with similarity-based aspects of determiner meaning, can also help in capturing those formal properties of determiners, such as monotonicity or definiteness, that theoretical semanticists have been traditionally interested in.

[7] The *observed* method is in fact at advantage in our experiment because a considerable number of DP foils are not found in the corpus and are assigned similarity 0 with the target.

7 Acknowledgments

This research was supported by the ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

References

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

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-Chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of EACL*, pages 23–32, Avignon, France.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. 2012. Distributional semantics in Technicolor. In *Proceedings of ACL*, pages 136–145, Jeju Island, Korea.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36:345–384.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Dan Garrette, Katrin Erk, and Ray Mooney. 2013. A formal approach to linking logical form and vector-space lexical semantics. In H. Bunt, J. Bos, and S. Pulman, editors, *Computing Meaning, Vol. 4*. In press.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of EMNLP*, pages 1394–1404, Edinburgh, UK.

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

Thomas Landauer and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

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

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

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

Richard Socher, Brody Huval, Christopher Manning, and Andrew Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP*, pages 1201–1211, Jeju Island, Korea.

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

Fabio Zanzotto, Ioannis Korkontzelos, Francesca Falucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of COLING*, pages 1263–1271, Beijing, China.