How Well Do Distributional Models Capture Different Types of Semantic Knowledge?

Dana Rubinstein **Effi Levi** **Roy Schwartz** **Ari Rappoport**
Institute of Computer Science, The Hebrew University
{drubin80,efle,roys02,arir}@cs.huji.ac.il

Abstract

In recent years, distributional models (DMs) have shown great success in representing lexical semantics. In this work we show that the extent to which DMs represent semantic knowledge is highly dependent on the type of knowledge. We pose the task of predicting properties of concrete nouns in a supervised setting, and compare between learning taxonomic properties (e.g., *animacy*) and attributive properties (e.g., *size*, *color*). We employ four state-of-the-art DMs as sources of feature representation for this task, and show that they all yield poor results when tested on attributive properties, achieving no more than an average F-score of 0.37 in the binary property prediction task, compared to 0.73 on taxonomic properties. Our results suggest that the distributional hypothesis may not be equally applicable to all types of semantic information.

1 Introduction

The Distributional Hypothesis states that the meaning of words can be inferred from their linguistic environment (Harris, 1954). This hypothesis lies at the heart of distributional models (DMs), which approximate the meaning of words by considering the statistics of their co-occurrence with other words in the lexicon.

DMs have shown impressive results in many semantic tasks, such as predicting the similarity of two words, grouping words into semantic categories, and solving analogy questions (see Baroni et al. (2014) for a recent survey). They are also used as a source of semantic information by many downstream applications, including syntactic parsing (Socher et al., 2013), image annotation (Klein et al., 2014), and semantic frame identification (Hermann et al., 2014).

However, the empirical success of DMs may not be uniform across the full range of semantic knowledge. It has been argued that DMs can never grasp the full meaning of words, as many aspects of meaning are grounded in the physical world (Andrews et al., 2009). This claim relies chiefly on cognitive theory (Louwerse, 2011), and is somewhat supported in empirical findings (Baroni and Lenci, 2008; Andrews et al., 2009). Moreover, a recent study by (Hill et al., 2014) has shown that DMs may not model word similarity as well as previously believed.

In this work, we seek to further study the capabilities of DMs in capturing semantic information. For our purposes, we assume that the meaning of a word referring to a *concrete object* (henceforth *concept*) is comprised of a list of *properties* (Baroni and Lenci, 2008). For example, the meaning of the concept *an apple* is comprised of such properties as *red, round, edible, a fruit*, etc. We distinguish between *taxonomic* properties (Wu and Barsalou, 2001; McRae et al., 2005), which define the conceptual category that a concept belongs to (e.g. *an apple is a fruit*), and all other types of properties (henceforth referred to as *attributive* properties). In this paper we employ DMs in the task of learning properties of concepts, and show a very large discrepancy in performance between learning taxonomic and attributive properties.

Several previous works addressed semantic property learning, but mostly in terms of automatically extracting salient properties of concepts from raw text (Almuhareb and Poesio, 2005; Barbu, 2008; Baroni and Lenci, 2008; Devereux et al., 2009; Baroni et al., 2010; Kelly, 2013). Baroni and Lenci (2008) is the only work we are aware of that addressed different property types, while utilizing a DM for property extraction. However, their approach is simple, and includes defining the properties of a concept to be the 10 neighboring words of that concept in the DM space.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 726–730,
Beijing, China, July 26-31, 2015.

In order to determine to what extent properties of concepts are captured by DMs, we define the following task. The goal is to predict, for a given concept, whether it holds a specific property or not (e.g., whether or not the concept *elephant* is considered *large*) . We model this task as a learning problem, in which concepts have a feature representation based on a state-of-the-art DM. A property-predictor is then trained to predict, for any given concept, whether the property applies to it or not (in a binary classification setup), or the strength of affiliation between the property and the concept (in a regression setup). By evaluating the performance of these predictors, we assess the degree to which the property is captured by the DM.

We experiment with four state-of-the-art DMs (Baroni and Lenci, 2010; Mikolov et al., 2013; Levy and Goldberg, 2014; Pennington et al., 2014). Our results show that all DMs, quite successful in many semantic tasks, fail when it comes to predicting attributive properties of concepts. For example, in the classification task, the best performing DM achieves an averaged F-score of only 0.37, contrasted with an average F-score of 0.73 achieved by the same model for taxonomic properties. This result, which may be attributed to an essential difference between taxonomic and attributive properties, demonstrates possible limitations of the distributional hypothesis, at least in terms of the information captured by current state-of-the-art DMs.

2 Learning Semantic Properties of Concepts

The goal of this paper is to gain better understanding of the type of information DMs encode. We do so by evaluating the performance of a predictor trained on a DM-based representation to learn a semantic property. In this section, we describe the proposed learning task, the dataset and the DMs which serve as feature representations.

2.1 Task Description

We model the problem of learning a single semantic property both as a binary classification problem and as a regression problem. The binary setup is simpler, however it may be argued that a regression setup is more appropriate, since the nature of the affiliation between a concept and its properties is not necessarily binary.

Binary Classification. For each property p, we take concepts for which p applies to be positive instances, and concepts for which it does not as negative instances. For example, the property *is loud* is positive for *a trumpet* but negative for *a mouse*. Let $\mathcal{X}$ denote the domain of concepts, and $\mathcal{Y}_p = \{\pm 1\}$ denote the binary label space. Then for each property p we learn a predictor $h_p : \psi(\mathcal{X}) \rightarrow \mathcal{Y}_p$, where $\psi(\mathcal{X}) \subseteq \mathbb{R}^n$ is a mapping from the concept domain to some DM space.

Regression. Here we consider the saliency of a property for a concept and regard it as a real-valued measure. For example, *white* is a salient property of *swan*, a less salient property of *house*, and not a property at all of *hammer*. The formal definitions are the same as in the binary classification setup, except that here $\mathcal{Y}_p = \mathbb{R}$.

2.2 The Data

We use the McRae Feature Norms dataset (McRae et al., 2005). This data was collected in a set of experiments, where participants were presented with concepts (concrete nouns only) and were asked to write down properties that describe them. This resulted in a matrix of 541 concepts and 2,526 properties, where each (concept, property) entry holds the number of participants who elicited the property for the concept. This dataset has been widely used in the past as a proxy to the human perceptual representation of concrete objects (Baroni and Lenci, 2008; Barbu, 2008; Devereux et al., 2009; Johns and Jones, 2012).

In the binary classification setting, for each property, we take all concepts for which this property was elicited (by any number of participants)[1] to be positive, and all other concepts to be negative. In the regression setting, we take the $[0, 1]$-scaled number of participants who elicited each property for a concept to be the real-valued measure of its saliency for that concept.

2.3 Distributional Models

We experiment with four state-of-the-art DMs as feature representations for the concept domain. The models differ with respect to their method of generation (neural network or transformed co-occurrence counts) and their consideration of lin-

[1] Due to a pre-defined threshold applied by McRae et al. (2005), only properties mentioned by at least 5 participants are considered positive.

guistic information (using plain text only, morphology, syntax or pattern information).

word2vec. word2vec (*w2v*, Mikolov et al. (2013)) is a neural network model which implements a language model objective. It has reached state-of-the-art results for word similarity, categorization and analogy tasks (Baroni et al., 2014). We use the off-the-shelf 300-dimensional version trained on a corpus of 100B tokens.[2]

GloVe. GloVe (*gv*, Pennington et al. (2014)) is a log bilinear regression model. The authors report state-of-the-art results in word similarity, semantic analogies and NER tasks. We use the off-the-shelf 300-dimensional version trained on a corpus of 840B tokens.[3]

Distributional Memory. The Distributional Memory model (*dm*, Baroni and Lenci (2010)) is a co-occurrence based DM, which admits morphological, structural and pattern information. The authors have shown that it is highly competitive with state-of-the-art co-occurrence models in a range of semantic tasks. We use the off-the-shelf 5K-dimensional version trained on 3B tokens.[4]

Dependency word2vec. The dependency word2vec model (*dep*, Levy and Goldberg (2014)) is a variation of the word2vec model, which takes into account the dependency links between words. The authors have shown that it accurately models word similarity. We use the off-the-shelf 300-dimensional version trained on Wikipedia.[5]

2.4 Experimental Setup

In our experiments, we consider properties which have at least 25 positive instances in the dataset. We then discard attributive properties that clearly correspond to a taxonomic property. For example, the property *has feathers* is no different from the *bird* category, or the property *lives in water* is identical to the *fish* category. The final list consists of 7 taxonomic and 13 attributive properties.[6]

For each property, we learn both a linear SVM classifier in the binary setup, and a linear SVM regressor in the regression setup. For both setups we use the lib-svm package (Chang and Lin, 2011)[7] and follow a 5-fold cross-validation protocol.

In the binary setup, we report F-scores only, as accuracy measures tend to be misleading due to an unbalanced label distribution. In the regression setup, we report Pearson's correlation scores between predicted values and gold standard values.

2.5 Results

Table 1 shows our results in the binary setup (left side) and in the regression setup (right side) for all models. We display average scores separately for taxonomic and attributive properties.

The results for the binary setup show a rather low performance on learning attributive properties, attaining an average F-score of no more than 0.37 (*dep* model). This is emphasized when compared to the average performance on taxonomic properties, which is 0.73 for *dep*, and can be as high as 0.78 (*w2v*). The regression setup shows a similar trend; the average correlation for attributive properties is at most 0.28 (*dep*), compared to 0.59 for taxonomic properties.

While linear Support Vectors are a well-established method for classification and regression, we have attempted the same experiments with several other methods, including K-Nearest-Neighbors and Decision Trees for classification, and simple Least Squares for regression. In all cases, the results were found to be inferior to the ones obtained by the Support Vectors, while maintaining the discrepancy in performance between taxonomic and attributive property learning.

3 Discussion

Our results show that there is a great difference between the performance of DMs when used to predict taxonomic and attributive properties. Concretely, four state-of-the-art DMs fail to predict attributive properties, implying that even if the property information is indicated in text, it is signaled very weakly, at least by means of linguistic regularities captured by current, state-of-the-art DMs.

Our findings are in line with previous work, such as (Baroni and Lenci, 2008), who demonstrated that taxonomic properties are more dominant in text compared to attributive properties. This suggests that the distributional hypothesis may not be equally applicable to all types of semantic information, and in particular, it may be

[2] code.google.com/p/word2vec/

[3] nlp.stanford.edu/projects/glove/

[4] clic.cimec.unitn.it/dm/

[5] levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/

[6] The average number of positive instances per property is 42 for taxonomic properties and 61 for attributive properties.

[7] www.csie.ntu.edu.tw/~cjlin/libsvm

	Property	Binary Classification				Regression			
		w2v	gv	dm	dep	w2v	gv	dm	dep
Taxonomic	a bird	0.83	0.86	0.78	0.71	0.63	0.63	0.39	0.57
	a fruit	0.86	0.8	0.72	0.6	0.66	0.69	0.57	0.55
	a mammal	0.71	0.69	0.65	0.73	0.47	0.44	0.46	0.41
	a vegetable	0.74	0.81	0.75	0.7	0.65	0.69	0.54	0.56
	a weapon	0.72	0.64	0.67	0.77	0.61	0.58	0.48	0.58
	an animal	0.8	0.77	0.74	0.82	0.79	0.73	0.51	0.78
	clothing	0.81	0.84	0.64	0.81	0.63	0.69	0.36	0.67
	Average	**0.78**	**0.77**	**0.71**	**0.73**	**0.63**	**0.64**	**0.47**	**0.59**
Attributive	of different colors	0.44	0.41	0.33	0.46	0.36	0.32	0.22	0.38
	is black	0.24	0.2	0.17	0.22	0.09	0.17	0.13	0.15
	is brown	0.28	0.23	0.29	0.33	0.25	0.25	0.16	0.27
	is green	0.4	0.4	0.45	0.44	0.28	0.24	0.28	0.39
	is white	0.19	0.22	0.11	0.2	0.06	0.1	0.06	0.15
	is yellow	0.21	0.14	0.15	0.21	0.12	0.15	0.12	0.23
	is large	0.4	0.41	0.42	0.44	0.39	0.34	0.38	0.33
	is small	0.43	0.4	0.43	0.48	0.29	0.21	0.25	0.31
	is long	0.31	0.24	0.31	0.36	0.24	0.03	0.14	0.27
	is round	0.29	0.3	0.29	0.43	0.22	0.15	0.24	0.28
	is loud	0.35	0.27	0.3	0.36	0.33	0.25	0.15	0.23
	is dangerous	0.45	0.47	0.49	0.5	0.32	0.3	0.25	0.41
	is fast	0.41	0.34	0.29	0.35	0.33	0.32	0.19	0.26
	Average	**0.34**	**0.31**	**0.31**	**0.37**	**0.25**	**0.22**	**0.2**	**0.28**

Table 1: Results for the Property Learning Task. On the left: F-scores for the binary classification task. On the right: Pearson correlation scores for the regression task.

limited with respect to attributive properties.

An interesting observation is found in the relative success of DMs in predicting taxonomic properties. This result, in line with past research, e.g. (Schwartz et al., 2014), may be explained by considering taxonomic properties as a rich aggregate of attributive properties (Baroni and Lenci, 2010). For example, animals usually have legs and mouths, they make sounds, they can be killed, etc. This is contrasted with attributive properties such as *is white*, whose members do not have much in common, other than the property itself. We therefore hypothesize that although attributive properties may be signaled very weakly in text, as our results indicate, their accumulation is sufficient to distinguish concepts that share most of them from concepts that do not.

To demonstrate this, we turned back to the McRae dataset. For each property, we observed the vector of its values across all concepts in the dataset. We then found its 5 nearest neighbors in terms of correlation, and computed the average correlation with these neighbors, denoted c. Next, we compared the averaged c value for taxonomic properties with that of attributive properties. Taxonomic properties show an average c value of 0.62, compared to 0.32 only for attributive properties. This supports our hypothesis that members of taxonomic properties are similar to each other in various aspects, while members of attributive properties are much less so. This finding may provide a partial explanation as to why taxonomic properties are more easily learned compared to attributive properties, as demonstrated in this paper.

To conclude, we have shown that in the context of learning semantic properties, state-of-the-art distributional models perform differently with respect to the type of property learned. Our results serve as a basis for establishing the limitations to the distributional hypothesis. As future work we propose to further investigate the nature of the distributional hypothesis in its manifestation as DMs, possibly by considering a more fine grained distinction between property types. For example, we intend to compare the performance between properties grounded in the physical world, like colors

or size, and more abstract properties such as *dangerous* or *cute*.

Acknowledgments

We would like to thank Roi Reichart for his careful reading and helpful comments. This research was funded (in part) by the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI).

References

Abdulrahman Almuhareb and Massimo Poesio. 2005. Concept learning and categorization from the web. In *Proc. of CogSci*.

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116(3):463.

Eduard Barbu. 2008. Combining methods to learn feature-norm-like concept descriptions. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 9–16.

Marco Baroni and Alessandro Lenci. 2008. Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1):55–88.

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

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2):222–254.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Barry Devereux, Nicholas Pilkington, Thierry Poibeau, and Anna Korhonen. 2009. Towards unrestricted, large-scale acquisition of feature-based conceptual representations from corpus data. *Research on Language and Computation*, 7(2-4):137–170.

Zellig S Harris. 1954. Distributional structure. *Word*.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proceedings of ACL*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*.

Brendan T Johns and Michael N Jones. 2012. Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, 4(1):103–120.

Colin Kelly. 2013. Automatic extraction of property norm-like data from large text corpora.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2014. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*.

Omer Levy and Yoav Goldberg. 2014. Dependencybased word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.

Max M Louwerse. 2011. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2):273–302.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP*.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2014. Minimally supervised classification to semantic categories using automatically acquired symmetric patterns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1612–1623, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. 2013. Parsing with compositional vector grammars. In *In Proceedings of the ACL conference*. Citeseer.

Ling-Ling Wu and Lawrence W Barsalou. 2001. Grounding concepts in perceptual simulation: I: Evidence from property generation. *Under review http://userwww. service. emory. edu/˜ barsalou*.