Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning

Li Lucy[1]
lucy3@stanford.edu

Jon Gauthier[1,2]
jon@gauthiers.net

[1]Stanford Symbolic Systems [2]Stanford NLP Group

Abstract

Distributional word representation methods exploit word co-occurrences to build compact vector encodings of words. While these representations enjoy widespread use in modern natural language processing, it is unclear whether they accurately encode all necessary facets of conceptual meaning. In this paper, we evaluate how well these representations can predict perceptual and conceptual features of concrete concepts, drawing on two semantic norm datasets sourced from human participants. We find that several standard word representations fail to encode many salient perceptual features of concepts, and show that these deficits correlate with word-word similarity prediction errors. Our analyses provide motivation for grounded and embodied language learning approaches, which may help to remedy these deficits.

1 Introduction

Distributional approaches to meaning representation have enabled a substantial amount of progress in natural language processing over the past years. They center around a classic insight from at least as early as Harris (1954); Firth (1957):

> You shall know a word by the company it keeps. (Firth, 1957, p. 11)

Popular distributional analysis methods which exploit this intuition such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) have been critical to the success of many recent large-scale natural language processing applications (e.g. Turney and Pantel, 2010; Turian et al., 2010; Collobert and Weston, 2008; Socher et al., 2013; Goldberg, 2016). These methods operationalize distributional meaning via tasks where words are optimized to predict words which co-occur with them in text corpora. These methods yield compact word representations — vectors in some high-dimensional space — which are optimized to solve these prediction tasks. These vector representations form the foundation of practically all modern deep learning models applied within natural language processing.

Despite the success of distributional representations in standard natural language processing tasks, a small but growing consensus within the artificial intelligence community suggests that these methods cannot be sufficient to induce adequate representations of words and concepts (Kiela et al., 2016; Gauthier and Mordatch, 2016; Lazaridou et al., 2015). These sorts of claims, which often draw on experimental evidence from cognitive science (see e.g. Barsalou, 2008), are used to back up arguments for multimodal learning (at the weakest) or complete embodiment (at the strongest). Kiela et al. (2016) claim the following:

> ...the best way for acquiring human-level semantics is to have machines learn through (physical) experience: if we want to teach a system the true meaning of "bumping into a wall," we simply have to bump it into walls repeatedly.

Discussions like the one above have an intuitive pull: certainly "bump" is best understood through a sense of touch, just as "loud" is best understood through a sense of sound. It seems inefficient — or perhaps just wrong — to learn these sorts of concepts from distributional evidence.

All project code available at github.com/lucy3/grounding-embeddings.

Proceedings of the First Workshop on Language Grounding for Robotics, pages 76–85,
Vancouver, Canada, July 30 - August 4, 2017.

Despite the intuitive pull, there is not much evidence from a computational perspective that grounded or multimodal learning actually earns us anything in terms of general meaning representation. Will our robots and chat-bots be worse off for not having physically bumped into walls before they hold discussions on wall-collisions? Will our representation of the concept *loud* somehow be faulty unless we explicitly associate it with certain decibel levels experienced in the real world? Before we proceed to embed our learning agents in multimodal games and robot-shells, it is important that we have some concrete idea of how grounding actually affects meaning.

This paper presents a thorough analysis of the contents of distributional word representations with respect to this question. Our results suggest that several common distributional word representations may indeed be deficient in the sort of grounded meaning necessary for language-enabled agents deployed in the real world.

2 Related work

This paper uses semantic norm datasets to evaluate the content of distributional word representations. Semantic norm datasets consist of concepts and norms concerning their perceptual and conceptual features, as provided by human participants. They are a popular resource within psychology and cognitive science as models of human concept representation, and have been used to explain psycholinguistic phenomena from semantic priming and interference (Vigliocco et al., 2004) to the structure of early word learning in child language acquisition (Hills et al., 2009). Andrews et al. (2009) show how "experiential" semantic norm information can be used to model human judgments of concept similarity. They show that this semantic norm data provides information distinct from the information found in basic word representations. Our work extends the findings of Andrews et al. to a larger semantic norm dataset and evaluates particular implications within natural language processing.

A small NLP literature has compared distributional representations with semantic norm datasets and other external resources. Rubinstein et al. (2015) confirm that word representations are especially effective at predicting taxonomic features versus attributive features. Collell and Moens (2016) find that word representations fail to predict many visual features of concepts, and show how representations from computer vision models can help improve these predictions. Several studies have used distributional representations to reconstruct aspects of these semantic norm datasets (Herbelot and Vecchi, 2015; Fagarasan et al., 2015; Erk, 2016).

The majority of the NLP work in this space has focused on the downstream task of augmenting word representations with novel grounded information, often evaluating on standard semantic similarity datasets (Agirre et al., 2009; Bruni et al., 2012; Faruqui et al., 2015; Bulat et al., 2016). Young et al. (2014) develop an alternative operationalization of denotational meaning using image captioning datasets, and demonstrate gains over distributional representations on textual similarity and entailment datasets.

This applied work has demonstrated that *something* worthwhile is indeed gained by augmenting distributional representations with some orthogonal grounded or multimodal information. We believe it is critical to analyze the original successes and failures of distributional representations in order to motivate this move to grounded meaning representation.

	# word tokens	# word types
GloVe (Common Crawl)	840B	2.2M
GloVe (Wiki+Gigaword)	6B	400K
word2vec	100B	3M

Table 1: Statistics of the corpora used to produce the distributional representations used in this paper.

3 Meaning representations

3.1 Distributional meaning

This paper examines representations produced by two popular unsupervised distributional methods. Table 1 shows the statistics of the corpora used to generate these vectors.

GloVe: GloVe (Pennington et al., 2014) estimates word representations w_i by using them to reconstruct a word-word co-occurrence matrix X collected from a large text corpus:

$$L = \sum_{i,j=1}^{V} f(X_{ij}) \left(w_i^T w_j + b_i + b_j - \log X_{ij} \right)^2 \quad (1)$$

Dataset	# concepts	# features	C/F	F/C
McRae	541	2526	2.87	13.41
CSLB	638	2725	3.78	16.13

Table 2: Semantic norm datasets used in this paper. The final two columns show the mean concepts per feature / features per concept.

here $f(X_{ij})$ is a weighting function on word pairs and b_i, b_j are learned per-word bias terms.

We use two pre-trained GloVe vector datasets: one trained on a concatenation of Wikipedia 2014 and Gigaword 5 (GloVe-WG), and another trained on a Common Crawl dump (GloVe-CC).[1]

word2vec: word2vec (Mikolov et al., 2013) estimates word representations by optimizing a skip-gram objective to predict all words w_j within a context window c of a word w_i given their word representations:

$$J = \frac{1}{T}\sum_{i=1}^{T}\sum_{i-c\leq j\leq i+c} \log p(w_j \mid w_i) \qquad (2)$$

where T is the total number of words in a corpus. We use a publicly available word2vec dataset trained on the Google News corpus.[2]

3.2 Semantic norms

Semantic feature norm datasets consist of reports from human participants about the semantic features of various natural kinds. A proportion of the features contained in these datasets are properties of concepts which may be obvious to humans but are perhaps difficult to find written in text corpora. For this reason, we selected two semantic norm datasets to serve as gold-standard comparisons of concept meaning. Table 2 displays basic statistics about the semantic norm datasets we use in this paper.

McRae Our initial experiments use the semantic norm dataset from McRae et al. (2005), which consists of 541 concrete noun concepts with associated feature norms, collected from 725 participants. For a given concept, the McRae dataset includes all feature norms which were reported independently by at least five participants (2,526 in total). After removing concepts indicated to have ambiguous meanings to mitigate polysemy effects (such as `tank_(army)` and `tank_(container)`) and one concept without a GloVe representation (`dunebuggy`), we had a resulting set of 515 concepts for analysis. The dataset groups features into several perceptual and non-perceptual categories: taxonomic, encyclopedic, function, visual-motion, visual-form_and_surface, visual-colour, sound, tactile, and taste (McRae et al., 2005). We use the McRae dataset and feature categories to perform basic pilot analyses and form hypotheses about the nature of the distributional representations tested.

CSLB We reproduce and extend our results on a second semantic norm dataset collected by the Cambridge Centre for Speech, Language and the Brain (CSLB; Devereux et al., 2014). CSLB contains 638 concepts provided by 123 participants. Their data collection closely followed McRae et al. (2005), though features were included if at least 2 participants named that feature. We removed concepts with two-word names, ambiguous meanings, or missing vector representations to yield a vocabulary of 597 concepts from this dataset. CSLB also includes a feature categorization schema, though the categories are broader than those in McRae: visual perceptual, other perceptual, functional, taxonomic, and encyclopedic.

The mapping between the two categorization schemes is far from perfect. While some perceptual features in McRae are categorized as perceptual features in CSLB, other features (e.g. those related to swimming, flying, eating) are reclassified as "functional" in CSLB. The two datasets disagree on abstract conceptual properties as well. For example, CSLB classifies `is_for_football` as a functional property, while McRae classifies the comparable feature `associated_with_football_games` as encyclopedic.

The encyclopedic category is somewhat difficult to distinguish in both datasets. It is composed mainly of abstract factual features, but also contains attributive features such as `is_cold_blooded` and `does_use_electricity` as well as `is_scary` and `is_cool`.

Meanwhile, the functional category mixes features for behaviors associated with the concept (`does_dive`) as well as functions that people perform on or with the concept (`is_hit`). This classification system may need some readjustments to provide a clear understanding of what is

[1] nlp.stanford.edu/projects/glove
[2] code.google.com/archive/p/word2vec

perceptual and what is conceptual, and it may be that some features, such as `has_a_steering_wheel`, are both.

Given the significant noise of this classification scheme, we focus our investigation on a single contrast between features in clearly perceptual categories (visual, tactile, sound, etc.) and non-perceptual categories (functional and taxonomic). Because the encyclopedic category contains an ambiguous mix of both sorts, we exclude it from our formal predictions later in the paper.

4 The feature view

We first investigate how well distributional word representations directly encode information about semantic norms.[3] For each feature in a semantic norm dataset, we construct a binary classification problem which predicts the presence or absence of the feature for each concept. Concretely, for each feature f_i we have a label vector $y_i \in \{0,1\}^{n_c}$, where n_c is the total number of concepts in the dataset, and y_{ij} is 1 when concept j has feature f_i and 0 otherwise. We build label vectors only for features with five or more associated concepts. After filtering, we have $n_f = 267$ label vectors in the McRae dataset and $n_f = 775$ in CSLB.

For each feature, we construct a binary logistic regression model p^i which predicts the presence or absence of the feature for a concept given its word representation x_j:

$$p^i(y_{ij} \mid x_j) = \sigma(w_i^T x_j) \tag{3}$$

This base model is extremely prone to overfitting, as most features have only several associated concepts — that is, each classifier has only a few positive examples — and the input word representations are of a high dimensionality. In order to prevent overfitting, we add an independent L2 regularization term to each regression model. For each feature f_i, we use leave-one-out cross-validation to select the regularization parameter λ_i which maximizes the following modified logistic objective:

$$L_i(\lambda_i) = \frac{1}{|f_i|} \sum_{x_j \in f_i} \left(\log p_{-j}^{i,\lambda_i}(y_{ij} = 1 \mid x_j) + \frac{1}{n_c - |f_i|} \sum_{x_k \notin f_i} \log p_{-j}^{i,\lambda_i}(y_{ik} = 0 \mid x_k) \right) \tag{4}$$

Here $p_{-j}^{i,\lambda_i}(\cdot)$ represents a regression model (Equation (3)) trained without example (x_j, y_{ij}) in the training set and with regularization parameter λ_i. The first term of the summand calculates the log-probability of the left-out concept having the desired feature, and the second term calculates the average log-probability that any other concept (outside of the feature group f_i) does not have the feature. The regularization terms λ_i are selected independently for each feature to maximize the objective L_i.

After fitting the regularized logistic regression models, we calculate a set of "feature fit" metrics. For each feature f_i, we evaluate the binary F1 score of its classifier's predictions $p^i(y_i)$. Figure 1 shows each feature as a point in a swarm-plot (grouped by feature category).

Pilot tests with the McRae dataset suggested that the categories associated with strictly perceptual features were not well encoded in the distributional representations relative to strictly non-perceptual categories (taxonomic and functional features).

We use the CSLB dataset as a test set for this prediction. We perform a bootstrap confidence interval test on the difference between the median feature fit scores for CSLB features in non-perceptual and perceptual categories. The 95% confidence intervals on this bootstrap are positive for two of the three representations tested (GloVe-CC and word2vec).[4] Figure 1 shows the feature fit scores on CSLB evaluated with GloVe-CC, and the word2vec evaluation effectively shows the same result: taxonomic and functional features score higher on average than strictly perceptual features. This comparison failed on GloVe-WG, however, where features classed as "functional" scored far lower on average than those in perceptual categories. Across all three sets of distributional representations, the median score of ency-

[3] The remainder of this paper describes a general analysis performed on both the McRae and CSLB datasets. We used McRae as a pilot dataset to form hypotheses, and checked these hypotheses on the CSLB dataset as a test set. All of the graphs and numbers reported in this paper correspond to results on CSLB.

[4] GloVe-CC: (7.67%, 24.0%); word2vec: (7.13%, 20.6%); GloVe-WG: (-1.25%, 15.7%).

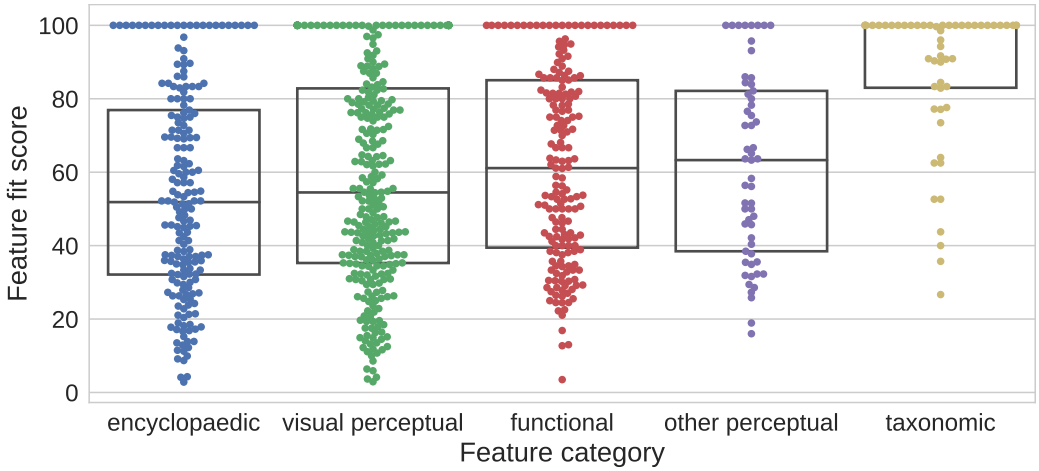


Figure 1: The CSLB feature fit metrics of GloVe-CC, where each point is a feature with at least 5 associated concepts. Feature categories are on the horizontal axis.

category	feature fit $< 50\%$	feature fit $> 50\%$
other perceptual	is_chewy, is_solid, is_high_pitched	is_hard, does_smell_good_nice, is_juicy
visual perceptual	is_triangular, has_a_string, is_curved,	has_a_clasp, has_a_shell, has_whiskers
encyclopedic	is_collectable, is_powerful, made_of_tissue	is_formal, does_not_fly, is_kept_in_a_cage
functional	is_roasted, is_for_weddings, is_carried	does_shelter, does_chop, is_eaten_edible
taxonomic	is_a_home, is_a_vessel, is_an_ingredient	is_seafood, is_a_boat, is_a_tool

Table 3: Examples of features in each category with feature fit scores based on using GloVe-CC to predict norms from CSLB.

clopedic features was well below all other feature categories.

It is obvious from Figure 1 that each category contains a wide range of feature fit values. As discussed earlier in Section 3.2, this categorization of features is far from perfect. Many of the lower-scoring features classed as "encyclopedic" are simple attributive features not deserving of the category label, such as `is_fresh` and `is_filling`. Many of the higher-scoring encyclopedic features seem genuinely encyclopedic, such as `is_found_on_farms`; other high-scoring features are arguably "functional," such as `does_grow_on_trees`. Many of the higher scoring visual perceptual features state structural part-whole relations, such as `has_legs` and `has_an_engine`.

Table 3 provides more examples of low- and high-scoring features in each category. Despite the rather noisy classification scheme used in this dataset, we still managed to find a regular trend in two of three evaluations, matching our expectations from prior pilot experiments. We believe that a revised classification scheme could help to demonstrate a clear difference between perceptual and non-perceptual features in all three datasets.

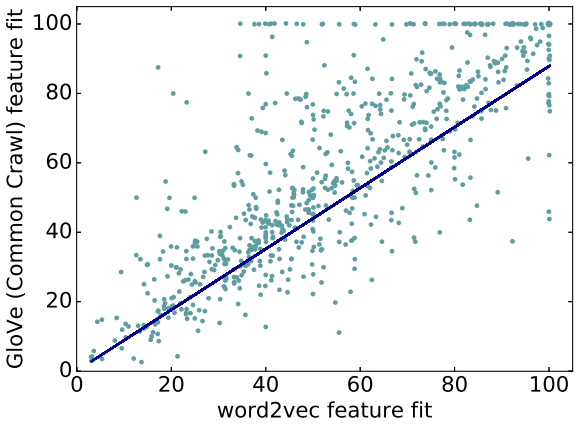
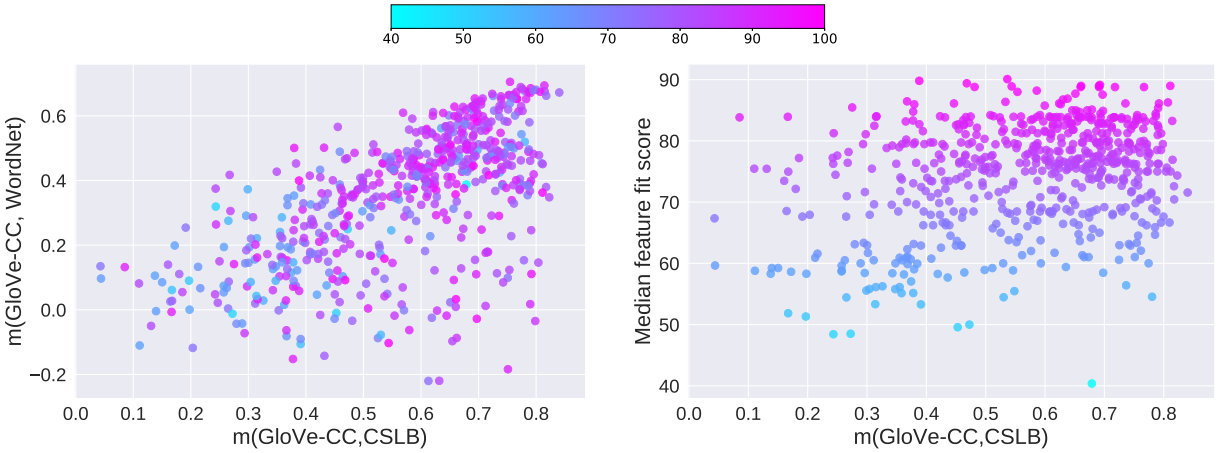


Figure 2: A comparison of CSLB feature fit scores for word2vec and GloVe-CC. Slope: 0.8773; Pearson r: 0.8260.

4.1 Matching word representation sources

For each feature, we compare its feature fit score evaluated with GloVe-CC word vectors and its score evaluated with word2vec vectors in Figure 2.



(a) Concepts plotted according to the correlation between their GloVe-CC embeddings and two other sources (CSLB, horizontal axis; WordNet, vertical axis). The points are colored according to their feature fit.

(b) A direct comparison of $m(\text{GloVe-CC}, \text{CSLB})$ (horizontal axis) and the median feature fit scores associated with concept (vertical axis). See main text for statistical tests.

Figure 3: Concept view results.

The trend in the figure suggests that both representations have similar feature fit deficiencies and strengths, though the trend becomes weaker near the $(100\%, 100\%)$ corner — the two representations correlate well at low feature fit scores, and seem to fan out at higher scores. A large group of points also sit in the figure at $y = 100$ and $x = 100$; these features are perfectly captured by one representation and not by the other.

This correlation is somewhat surprising, given that the word2vec and GloVe vectors are the products of different algorithms executed on very different corpora. There are two likely explanations behind this correlation:

1. Some features in the CSLB semantic norm data are unusually difficult, or are perhaps missing associated concepts. GloVe and word2vec correlate in performance because they don't match these noisy or incomplete features.

2. There are systematic deficiencies in the word vectors due to their shared reliance on the distributional method.

It is difficult to differentiate these two explanations on these small semantic norm datasets, but we hope to distinguish these in the future by testing new predictions for concepts not covered in these datasets. We will return to this idea in the conclusion of the paper.

5 The concept view

The previous section demonstrated that several classes of perceptual features are not well encoded on average by distributional word representations, and that these deficiencies systematically match across representations. How does this deficiency in feature representation carry over into computations on the word representations themselves?

We evaluate the matching between distributional representations and representations from other sources by comparing their predictions of word-word similarity. For distributional word representations, we compute word-word similarity by cosine distance:

$$\text{sim}(i, j) = \cos(x_i, x_j) \tag{5}$$

We derive compact concept representations from the semantic norm datasets with LSA (Landauer et al., 1998). We compute a truncated SVD on the feature matrix $Y \in \{0, 1\}^{n_c \times n_f}$, which is the concatenation of the binary feature label vectors introduced in Section 4. We define concept-concept similarity by the cosine distance between their corresponding LSA vectors.

As a secondary data source, we also compute word-word similarity judgments from the WordNet taxonomy (Miller, 1995). We use the Resnik metric (Resnik et al., 1999) to compute the similarity between concept names c_i, c_j:

$$\text{sim}_{\text{resnik}}(c_i, c_j) = \max_{c \in S(c_i, c_j)} -\log p(c) \tag{6}$$

where $S(c_i, c_j)$ selects the common ancestors of the concepts in the WordNet taxonomy, and $p(c)$ is the unigram probability of a concept as computed on an external corpus. This selects the ancestor of the two concepts in the taxonomy which has maximal information content (surprisal). We use WordNet as additional verification that the trends observed between semantic norms and distributional representations are non-coincidental.

We use these similarity metrics to compute pairwise distance measures for concepts present in the semantic norm datasets. For each metric, we produce a symmetric pairwise distance matrix $D \in \mathbb{R}^{n_c \times n_c}$, where an element D_{ij} indicates the distance between concepts i and j according to the metric.

We next compute how well each concept's pairwise similarity is correlated between the various metrics. For a given concept, we compute the Pearson correlation between the concept's GloVe/word2vec pairwise distance vector and the LSA and WordNet pairwise distance vectors.[5] The correlation values of interest are $m(\text{GloVe/word2vec}, \text{CSLB})$ and $m(\text{GloVe/word2vec}, \text{WordNet})$ — that is, the correlations between the pairwise distance vectors for GloVe/word2vec and CSLB and between the pairwise distance vectors for GloVe/word2vec and WordNet.

Figure 3a plots both of these correlation values evaluated with GloVe-CC for all concepts. The two m measures are evidently positively correlated, though with some noise ($r = 0.6160$). This is to be expected, as the CSLB dataset and WordNet overlap only partially in the semantic features they encode.

Each concept in Figure 3 is colored according to the median feature fit score of its associated features. In Figure 3b, we show this feature fit metric on the vertical axis. There is a positive relationship here between feature fit scores and the correlation metric $m(\text{GloVe-CC}, \text{CSLB})$ ($r = 0.3323$). Because the correlation between $m(\cdot, \text{CSLB})$ and feature fit metrics is weaker than expected, we run post-hoc multiple regression significance tests for each distributional representation. An F-test shows that the regression feature $m(\cdot, \text{CSLB})$ significantly improves predictions of feature fit values relative to a baseline model for all three representations[6,7].

[5]The Pearson correlation between two vectors is equivalent to the cosine distance between their mean-centered forms.

Domain	Feature fit	Concepts
10	61.03%	bread, cheese, chocolate, coffee, glue, ham, jam, jelly, ketchup, moss, soup, tea, yoghurt
15	67.18%	artichoke, asparagus, aubergine, bean, cabbage, flour, gherkin, leek, mango, pineapple, potato, pumpkin, rhubarb, seaweed
25	75.62%	bouquet, buttercup, carnation, daffodil, daisy, dandelion, fern, geranium, hyacinth, lily, marigold, orchid, pansy, poppy, rose, sunflower, tulip
31	78.22%	bayonet, bomb, cannon, crossbow, dagger, grenade, gun, pistol, revolver, rifle, shotgun, sword
36	82.18%	book, catalogue, menu, dictionary, encyclopaedia, textbook

Table 4: Selected domains from the clustering analysis on GloVe-CC, with median feature fit scores over concepts.

There is substantial variance in the predictions of the distributional representations due to factors outside of the scope of the semantic norm data. The mismatch in predictions between distributional representations is nevertheless a statistically significant predictor of feature fit metrics. This suggests that the feature-level deficiencies discovered in the previous section have concrete implications in terms of word-word similarity measures.

5.1 Domain-level analysis

We next investigate whether some domains of concepts are particularly affected by the deficiencies discussed in the previous sections. We perform agglomerative clustering on concepts from the CSLB dataset using a custom distance metric:

$$d(i,j) = ||\text{LSA}_i - \text{LSA}_j||_2 + \alpha(\text{FF}_i - \text{FF}_j)^2 \quad (7)$$

where LSA_i is the LSA vector representation computed from the semantic norm data for concept i as introduced earlier in this section, and FF_i is the median feature-fit score for a concept i. We select the weight α manually to produce the most semantically coherent clusters.

[6]The baseline regression model predicts a concept's feature fit from these baseline features: log(word frequency in Brown corpus), log(# associated features), log(total # feature reports for the concept), # WordNet senses.

[7]GloVe-CC: $F^* = 41.297, p < 10^{-9}$; GloVe-WG: $F^* = 68.783, p < 10^{-15}$, word2vec: $F^* = 41.27, p < 10^{-9}$

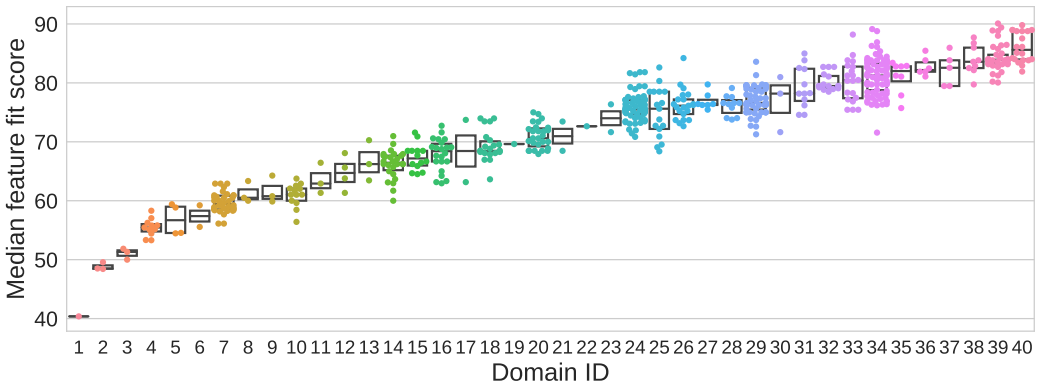


Figure 4: Concept domains derived from the CSLB semantic norm data. Each point represents a concept. The vertical axis is the median feature fit score of the concept's features on GloVe-CC.

Figure 4 shows the distribution of feature fit scores for each of the resulting 40 domains. We find that settings of α which yield semantically coherent clusters also yield groups of concepts with very low variance in feature fit scores. In Table 4 we list select domains and their median feature fit scores. This clustering suggests that deficiencies at the feature level affect entire coherent semantic domains of concepts.

6 Conclusion

This paper has analyzed how well various standard distributional representations encode aspects of grounded meaning. We chose to use semantic norm datasets as a gold standard of grounded meaning, and tested how word representations predicted features within these datasets. We grouped these features into high-level categories and found that, despite large within-category variance, several standard distributional representations underperformed on average in predicting perceptual features. The difference in prediction performance proved statistically significant on two of the three representations we evaluated. These deficiencies in feature encoding matched between GloVe and word2vec representations trained on different corpora, suggesting that certain classes of features may be poorly represented by distributional methods in general.

We also examined the consequences of these deficiencies in feature encoding for the word representations themselves. We compared the word-word similarity predictions made with distributional representations with those made with the semantic norm dataset and with WordNet, and found that words having features badly encoded within the distributional representations were also likely to make different similarity predictions than the predictions from these two corpora. A final domain-level concept analysis suggested that some semantic domains are particularly impacted by these issues in feature encoding.

The semantic norm datasets used in this paper are subject to saliency biases: they only contain the concept-feature mappings which experimental subjects think to mention when queried. These saliency effects add noise to our results, as mentioned in Section 4.1, and may have caused us to generally underestimate the performance of distributional models within all feature categories. In future work, we plan to repeat the sorts of tests conducted in this paper while avoiding possible saliency confounds. We also plan to develop a causal explanation for the deficiencies in the word embeddings found in this paper, showing how co-occurrence information (or lack thereof) present in the training corpus can bias performance on these tasks. Both of these studies will verify that the results we have found are due entirely to deficiencies in distributional methods rather than in the datasets used here.

We think these deficiencies should be worrying: if neural models of language are to have any knowledge about concepts, it ought to be in their word embeddings. Our findings show that these embeddings are lacking in basic features of perceptual meaning. These results suggest that distributional meaning (as operationalized by modern distributional models) may miss out on fundamental elements of semantics. We hope they will help motivate further work in developing multimodal representations which can prepare us to deploy more fluent language agents in the real world.

Acknowledgements

We thank Christopher D. Manning, Peng Qi, Pakapol Supaniratisai, Keenon Werling, and members of the Stanford, University of Washington, and Berkeley NLP communities for useful discussions, and the anonymous reviewers for their insightful comments.

References

Eneko Agirre, Enrique Alfonseca, Keith B. Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *HLT-NAACL*.

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

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.* 59:617–645.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional Semantics in Technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '12, pages 136–145. http://dl.acm.org/citation.cfm?id=2390524.2390544.

Luana Bulat, Douwe Kiela, and Stephen Clark. 2016. Vision and feature norms: Improving automatic feature norm learning through cross-modal maps. In *HLT-NAACL*.

Guillem Collell and Marie-Francine Moens. 2016. Is an image worth more than a thousand words? on the fine-grain semantic differences between visual and linguistic representations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 2807–2817. http://aclweb.org/anthology/C16-1264.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, pages 160–167.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods* 46(4):1119–1127.

Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics* 9(17):1–63. https://doi.org/10.3765/sp.9.17.

Luana Fagarasan, Eva Maria Vecchi, and Stephen Clark. 2015. From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *IWCS*.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. 2015. Retrofitting Word Vectors to Semantic Lexicons .

John Rupert Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis* .

Jon Gauthier and Igor Mordatch. 2016. A paradigm for situated and goal-driven language learning. *arXiv preprint arXiv:1610.03585* .

Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* 57:345–420.

Zellig S Harris. 1954. Distributional structure. *Word* 10(2-3):146–162.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: mapping distributional to model-theoretic semantic spaces. In *EMNLP*.

Thomas T Hills, Mounir Maouene, Josita Maouene, Adam Sheya, and Linda Smith. 2009. Categorical structure among shared features in networks of early-learned nouns. *Cognition* 112(3):381–396.

Douwe Kiela, Luana Bulat, Anita L Vero, and Stephen Clark. 2016. Virtual embodiment: A scalable long-term strategy for artificial intelligence research. *arXiv preprint arXiv:1610.07432* .

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes* 25(2-3):259–284.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *HLT-NAACL*.

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.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Philip Resnik et al. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)* 11:95–130.

Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. 2015. How Well Do Distributional Models Capture Different Types of Semantic Knowledge? In *ACL*.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*. Citeseer, volume 1631, page 1642.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, pages 384–394.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37:141–188.

Gabriella Vigliocco, David P Vinson, William Lewis, and Merrill F Garrett. 2004. Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive psychology* 48(4):422–488.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2:67–78.