From distributional semantics to feature norms: grounding semantic models in human perceptual data

Luana Făgărășan **Eva Maria Vecchi** **Stephen Clark**

Computer Laboratory
University of Cambridge
first.last@cl.cam.ac.uk

Abstract

Multimodal semantic models attempt to ground distributional semantics through the integration of visual or perceptual information. Feature norms provide useful insight into human concept acquisition, but cannot be used to ground large-scale semantics because they are expensive to produce. We present an automatic method for predicting feature norms for new concepts by learning a mapping from a text-based distributional semantic space to a space built using feature norms. Our experimental results are promising, and show that we are able to generalise feature-based representations for new concepts. This work opens up the possibility of developing large-scale semantic models grounded in a proxy for human perceptual data.

Classical distributional semantic models [1, 2] represent the meanings of words by relying on their statistical distribution in text [3, 4, 5, 6]. Despite performing well in a wide range of semantic tasks, a common criticism is that by only representing meaning through linguistic input these models are not grounded in perception, since the words only exist in relation to each other and are not grounded in the physical world. This concern is motivated by the increasing evidence in the cognitive science literature that the semantics of words is derived not only from our exposure to the language, but also through our interactions with the world. One way to overcome this issue would be to include perceptual information in the semantic models [7]. It has already been shown, for example, that models that learn from both visual and linguistic input improve performance on a variety of tasks such as word association or semantic similarity [8].

However, the visual modality alone cannot capture all perceptual information that humans possess. A more cognitively sound representation of human intuitions in relation to particular concepts is given by semantic property norms, also known as semantic feature norms. A number of property norming studies [9, 10, 11] have focused on collecting feature norms for various concepts in order to allow for empirical testing of psychological semantic theories. In these studies humans are asked to identify, for a given concept, its most important attributes. For example, given the concept AIRPLANE, one might say that its most important features are `to_fly`, `has_wings` and `is_used_for_transport`. These datasets provide a valuable insight into human concept representation and have been successfully used for tasks such as text simplification for limited vocabulary groups, personality modelling and metaphor processing, as well as a proxy for modelling perceptual information [12, 13].

Despite their advantages, semantic feature norms are not widely used in computational linguistics, mainly because they are expensive to produce and have only been collected for small sets of words; moreover the set of features that one can produce for a given concept is not restricted. In [14], the authors construct a three-way multimodal model, integrating textual, feature and visual modalities. However, this method is restricted to the same disadvantages of feature norm datasets. There have been some attempts at automatically generating feature-norm-like semantic representations for

concepts using large text corpora [15, 16, 17] but the generated features are often a production of carefully crafted rules and statistical distribution of words in text rather than a proxy for human conceptual knowledge. Our work focuses on automatic prediction of features for new concepts by learning a mapping from a distributional semantic space based solely on linguistic input to a more cognitively-sound semantic space where feature norms are seen as a proxy for perceptual information.

1 Mapping between semantic spaces

The integration of perceptual and linguistic information is supported by a large body of work in the cognitive science literature [12, 13] that shows that models that include both types of information perform better at fitting human semantic data. The idea of learning a mapping between semantic spaces appears in previous work; for example [18] learn a cross-modal mapping between text and images and [19] show that a linear mapping between vector spaces of different languages can be learned by only relying on a small amount of bilingual information from which missing dictionary entries can be inferred. Following the approach in [19], we learn a linear mapping between the distributional space and the feature-based space[1].

1.1 Feature norm datasets

One of the largest and most widely used feature-norm datasets is the one compiled by McRae and colleagues [9]. Participants were asked to produce a list of features for a given concept, whilst being encouraged to write down different kinds of properties, *e.g.* how the concept feels, smells or for what it is used (Table 1). The published dataset contains a total of 2526 features for 541 concrete concepts, with a mean of 13.7 features per concept.

SHRIMP	CUCUMBER
is_edible, 19	a_vegetable, 25
is_small, 17	eaten_in_salads, 24
lives_in_water, 12	is_green, 23
is_pink, 11	is_long, 15
tastes_good, 9	eaten_as_pickles, 12

Table 1: Examples of features and production frequencies for concepts from the McRae norms

More recently, the Cambridge Centre for Speech, Language and the Brain [11] collected semantic properties for 638 concrete concepts in a fashion similar to [9]. This is the largest publicly available feature norm dataset, containing 4359 features for 638 concepts (out of which 415 overlap with the McRae dataset), with an average of 2.15 features per concept more than in the McRae norms. There are also other property norms datasets which contain (besides concrete concepts) verbs and nouns referring to events [10]. Since the semantic property norms in the McRae dataset have been used extensively in the literature as a proxy for perceptual information, we will report our experimental results on this dataset.

1.2 Semantic spaces

A feature-based semantic space (**FS**) can be built using a similar architecture to the one used in co-occurrence based distributional models. Concepts are treated as target words, features as context words and co-occurrence counts are replaced with production frequencies (the number of participants that produced the feature for a given concept) (Table 2). We build two such feature-based semantic spaces: one using all the 2526 features in the McRae dataset as contexts (FS1) and one obtained by reducing the dimensions of the first space to 300 using SVD (FS2)[2].

[1] No word sense disambiguation was performed.

[2] All semantic spaces, both feature-based and distributional, were built using the DISSECT toolkit [20].

	has_fur	has_wheels	an_animal	a_pet	a_weapon
cat_FS	22	0	21	17	0
	dog	black	book	animal	breed
cat_DS	4516	3124	1500	2480	1631

Table 2: Example representation of CAT in the feature-based and distributional spaces

For the distributional spaces (**DS**), we experimented with various parameter settings for building co-occurrence based semantic spaces using Wikipedia as a corpus. We built a total of four models using the following parameters:

- DS1: contexts are the top 10K most frequent content words in Wikipedia, counts are raw co-occurrence counts between target and context words.
- DS2: same contexts as DS1, counts are re-weighted using PPMI and row normalised [21].
- DS3: perform SVD to 300 dimensions on DS2.
- DS4: same as DS3 but with row normalisation performed after dimensionality reduction

We also use the context-predicting type vectors available as part of the word2vec[3] project [5] (DS5). These vectors are 300 dimensional and are pre-trained on a Google News dataset (100 billion words).

1.3 The mapping function

Our goal is to learn a function f: **DS** $\rightarrow$ **FS** that maps a distributional vector for a concept to its feature-based vector. Following the hypothesis that many similarities amongst words can be represented as linear transformations [22], we learn the mapping as a linear relationship between the distributional representation of a word and its featural representation. We estimate the coefficients of the function using (multivariate) partial least squares regression (PLSR) as implemented in the R pls package [23], with the latent dimension parameter of PLSR set to 50.

2 Experimental results

We performed all experiments using a training set of 400 McRae randomly selected concepts and a testset of the remaining 138[4]. We use the featural representations of the words in the training set in order to learn a mapping between the two spaces, and the featural representations of the concepts in the test set as a gold-standard vector in order to analyse the quality of the learned transformation.

We performed a quantitative analysis as follows: for each item in the testset ($\vec{x}$), we computed the predicted[5] vector for the concept as $\overrightarrow{pred\_x} = f(\vec{x})$ followed by a retrieval of the top neighbours of $\overrightarrow{pred\_x}$ (using cosine similarity) in the feature-based semantic space. We were interested in observing, for a given concept, whether the gold-standard featural vector was retrieved in the topN neighbours of the predicted featural vector for that concept. Results averaged over the entire test set are summarised in Table 3.

A qualitative evaluation of the top neighbours for predicted featural vectors of concepts can be found in Table 4. Overall, the mapping results look promising, even for items that do not list the gold feature vector as one of the top neighbours. The neighborhoods of the predicted feature vectors correspond to those of the holistic vectors. However, overall the mapping looks too coarse. One reason could be the fact that the feature based space is relatively sparse (the maximum number of features for a concept is 26, whereas there are over 2500 dimensions in the space). The reason why, for example, the predicted vector for JAR does not contain its gold standard in the top 20 neighbours might simply be that there are not enough discriminating features for the model to learn that a jar usually has a lid and a bucket does not.

[3] https://code.google.com/p/word2vec/

[4] Out of the 541 McRae concepts, we discarded three (AXE, ARMOUR and DUNEBUGGY) because they were not available in the pre-trained word2vec vectors.

[5] vector in **FS** resulting from the mapping

Distributional space	Featural space	(%) in top1	(%) in top5	(%) in top10	(%) in top20
DS5	FS2	1.45	19.57	26.09	46.38
DS5	FS1	1.45	14.49	24.64	44.20
DS4	FS2	3.62	15.22	25.36	49.28
DS1	FS1	0.72	14.49	29.71	49.28
DS2	FS1	2.90	12.32	23.91	47.10
DS3	FS2	2.17	15.22	26.09	50.00

Table 3: Retrieval of gold-standard vectors in topN neighbours

Word	**Nearest neighbours**	Result
JAR	bucket, strainer, spatula, pot, cap_bottle	Not in top 20
JEANS	shawl, blouse, shirt, dress, sweater	Not in top 20
BUGGY	skateboard, scooter, truck, trolley, cart	In top20
SEAWEED	shrimp, perch, minnow, trout, squid	In top20
HORSE	donkey, cow, ox, sheep, goat	In top10
PISTOL	gun, rifle, revolver, shotgun, harpoon	In top10
SPARROW	starling, nightingale, finch, partridge, sparrow	In top5
SPATULA	tongs, spatula, strainer, colander, grater	In top5
HATCHET	hatchet, machete, sword, dagger, chisel	In top1
RAVEN	raven, chickadee, sparrow, falcon, partridge	In top1

Table 4: Example of nearest neighbours of predicted vectors for 10 concepts in the test set

3 Conclusion

Feature norms have shown to be potentially useful as a proxy for human conceptual knowledge and grounding, an idea that has been the basis of numerous studies psychological despite the limited availability of large-scale data for various semantic tasks. In this paper, we present a methodology to automatically predict feature norms for new concepts by mapping the representation of the concept from a distributional space to its feature-based semantic representation.

Clearly much experimental work is yet to be done, but in this initial study we have demonstrated the promise of such a mapping. We see two major advantages to our approach. First, we are no longer need limited to the sparse datasets and expensive procedures when working with feature norms, and second, we can gain a better understanding of the relationship between the distributional use of a word and our cognitive and experiential representation of the corresponding concept. We envisage a future in which a more sophisticated computational model of semantics, integrating text, vision, audio, perception and experience, will encompass our full intuition of a concept's meaning.

In future work, we plan to pursue this research in a number of ways. First, we aim to investigate ways to improve the mapping between spaces by exploring different machine learning approaches, such as other types of linear regression or canonical-correlation analysis. We are also interested in comparing the performance of non-linear transformations such as neural network embeddings with that of linear mappings. In addition, we wish to perform a more qualitative investigation of which distributional dimensions influence which feature norms in feature space. Lastly, we plan to evaluate **FS** representations on datasets that capture lexical similarity, such as the MEN dataset [8].

4 Acknowledgements

Luana Făgărășan is supported by an EPSRC Doctoral Training Grant. Eva Maria Vecchi is supported by ERC Starting Grant DisCoTex (306920). Stephen Clark is supported by EPSRC grant EP/I037512/1 and ERC Starting Grant DisCoTex (306920). We thank the anonymous reviewers for their helpful comments.

References

[1] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188, 2010.

[2] Magnus Sahlgren. *The Word-Space Model*. Dissertation, Stockholm University, 2006.

[3] Katrin Erk. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653, 2012.

[4] Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] Stephen Clark. Vector space models of lexical meaning. *Handbook of Contemporary Semantics–second edition. Wiley-Blackwell*, 2012.

[7] Lawrence W Barsalou, W Kyle Simmons, Aron K Barbey, and Christine D Wilson. Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2):84–91, 2003.

[8] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal Artificial Intelligence Research (JAIR)*, 49:1–47, 2014.

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

[10] David P Vinson and Gabriella Vigliocco. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1):183–190, 2008.

[11] Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, pages 1–9, 2013.

[12] Brian Riordan and Michael N Jones. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):303–345, 2011.

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

[14] Stephen Roller and Sabine Schulte im Walde. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[15] Colin Kelly, Barry Devereux, and Anna Korhonen. Automatic extraction of property norm-like data from large text corpora. *Cognitive Science*, 38(4):638–682, 2014.

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

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

[18] Angeliki Lazaridou, Elia Bruni, and Marco Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[19] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[20] Georgiana Dinu, Nghia The Pham, and Marco Baroni. DISSECT: DIStributional SEmantics Composition Toolkit. In *Proceedings of the System Demonstrations of ACL 2013*, East Stroudsburg, PA, 2013.

[21] Tamara Polajnar and Stephen Clark. Improving distributional semantic vectors through context selection and normalisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 230–238, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[22] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751. Citeseer, 2013.

[23] Björn-Helge Mevik and Ron Wehrens. The pls package: principal component and partial least squares regression in r. *Journal of Statistical Software*, 18(2):1–24, 2007.