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

Luana Făgărășan, **Eva Maria Vecchi** and **Stephen Clark**
University of Cambridge
Computer Laboratory
{luana.fagarasan|eva.vecchi|stephen.clark}@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 show that we are able to generalise feature-based concept representations, which opens up the possibility of developing large-scale semantic models grounded in a proxy for human perceptual data.

1 Introduction

Distributional semantic models (Turney and Pantel, 2010; Sahlgren, 2006) represent the meanings of words by relying on their statistical distribution in text (Erk, 2012; Bengio et al., 2006; Mikolov et al., 2013; Clark, 2015). 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 in relation to 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 (Barsalou et al., 2003). 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 (Bruni et al., 2014).

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 (McRae et al., 2005; Vinson and Vigliocco, 2008; Devereux et al., 2013) 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 the most important attributes of a concept; *e.g.* given AIRPLANE, its most important features could be `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 (Riordan and Jones, 2011; Andrews et al., 2009; Hill et al., 2014). Feature norms provide an interesting source of semantic information because they capture higher level conceptual knowledge in comparison to the low level perceptual information represented in images, for example.

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 there is no finite list of features that can be produced for a given concept. In Roller and Schulte im

Proceedings of the 11th International Conference on Computational Semantics, pages 52–57,
London, UK, April 15-17 2015.

SHRIMP	CUCUMBER	DRESS
is_edible, 19	a_vegetable, 25	clothing, 21
is_small, 17	eaten_in_salads, 24	worn_by_women, 15
lives_in_water, 12	is_green, 23	is_feminine, 10
is_pink, 11	is_long, 15	is_formal, 10
tastes_good, 9	eaten_as_pickles, 12	is_long, 10
has_a_shell, 8	has_skin, 9	different_styles, 9
lives_in_oceans, 8	grows_in_gardens, 7	made_of_material, 9

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

Walde (2013), 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 norms using large text corpora (Kelly et al., 2014; Baroni et al., 2010; Barbu, 2008) 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 predicting 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. A precedent for this work has been set in Johns and Jones (2012), but whilst they predict feature representations through global lexical similarity, we infer them through learning a cross-modal mapping.

2 Mapping between semantic spaces

The integration of perceptual and linguistic information is supported by a large body of work in the cognitive science literature (Riordan and Jones, 2011; Andrews et al., 2009) 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 Lazaridou et al. (2014) learn a cross-modal mapping between text and images and Mikolov et al. (2013) 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 Mikolov et al. (2013), we learn a linear mapping between the distributional space and the feature-based space.

2.1 Feature norm datasets

One of the largest and most widely used feature-norm datasets is from McRae et al. (2005). 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 dataset contains a total of 2526 features for 541 concrete concepts, with a mean of 13.7 features per concept. More recently, Devereux et al. (2013) collected semantic properties for 638 concrete concepts in a similar fashion. There are also other property norms datasets which contain verbs and nouns referring to events (Vinson and Vigliocco, 2008). 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.

2.2 Semantic spaces

A feature-based semantic space (**FS**) can be represented in a similar way to the co-occurrence based distributional models. Concepts are treated as target words, features as context words and co-occurrence counts are replaced with production frequencies, *i.e.* 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

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

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

features in the McRae dataset as contexts (FS1) and one obtained by reducing the dimensions of FS1 to 300 using SVD (FS2).

For the distributional spaces (**DS**), we experimented with various parameter settings, and built four spaces using Wikipedia as a corpus and sentence-like windows together with the following parameters:

- DS1: contexts are the top 10K most frequent content words in Wikipedia, values are raw co-occurrence counts.
- DS2: same contexts as DS1, counts are re-weighted using PPMI and normalised as detailed in Polajnar and Clark (2014).
- 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 vectors available as part of the word2vec[1] project (Mikolov et al., 2013) (DS5). These vectors are 300 dimensional and are trained on a Google News dataset (100 billion words).

2.3 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 Mikolov et al. (2013), 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 (Mevik and Wehrens, 2007), with the latent dimension parameter of PLSR set to 50.

3 Experimental results

We performed all experiments using a training set of 400 randomly selected McRae concepts and a test set of the remaining 138.[2] We use the featural representations of the concepts 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 gold-standard vectors in order to analyse the quality of the learned transformation.

For each item in the test set, we computed the concept's predicted vector, $f(\vec{x})$, by applying the learned mapping, f, to the concept's representation in **DS**, $\vec{x}$. We then retrieved the top neighbours of the predicted vector in **FS** using cosine similarity. 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. Results averaged over the entire test set are summarised in Table 3. We also report the performance of a random baseline (RAND), where a concept's nearest neighbours are randomly ranked, and we note that our model outperforms chance by a large margin.

For the experiments in which the feature space dimensions are interpretable, *i.e.* not reduced (FS1), we also report the MAP (Mean Average Precision). This allows us to measure the learnt mapping's ability to assign higher values to the gold features of a McRae concept (those properties that have a non-zero production frequency for a particular concept in the McRae dataset) than to the non-gold features.

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

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

DS	FS	**top1**	**top5**	**top10**	**top20**	**MAP**
RAND	-	0.37	0.74	1.85	3.70	-
DS1	FS1	0.72	14.49	29.71	49.28	0.30
DS2	FS1	2.90	12.32	23.91	47.10	0.29
DS3	FS1	2.90	13.04	24.64	49.28	0.37
DS3	FS2	2.17	15.22	26.09	50.00	-
DS4	FS2	3.62	15.22	25.36	49.28	-
DS5	FS1	1.45	14.49	24.64	44.20	0.29
DS5	FS2	1.45	19.57	26.09	46.38	-

Table 3: Percentage (%) of test items that retrieve their gold-standard vector in the topN neighbours of their predicted vector.

Word	**Nearest neighbours of predicted vector**	**Result**	**Top weighted predicted features**
JAR	bucket, strainer, pot, spatula	not top20	made_of_plastic, is_round*, made_of_metal, found_in_kitchens*
JEANS	shawl, shirt, blouse, sweater	not top20	clothing, different_colours, worn_by_women*
BUGGY	skateboard, truck, scooter, cart	in top20	has_wheels, made_of_wood*, is_large*, used_for_transportation
SEAWEED	shrimp, perch, trout, salmon	in top20	is_edible, lives_in_water*, is_green, swims*, is_small*
HORSE	cow, ox, sheep, donkey	in top10	an_animal, has_4_legs, is_large, has_legs, lives_on_farms
PLATYPUS	otter, salamander, turtle, walrus	in top10	an_animal, is_small*, lives_in_water, is_long*,
SPARROW	starling, finch, partridge, sparrow	in top5	a_bird, flies, has_feathers, has_a_beak, has_wings
SPATULA	strainer, spatula, grater, colander	in top5	made_of_metal, found_in_kitchens, made_of_plastic
HATCHET	hatchet, machete, sword, dagger	in top1	made_of_metal, is_sharp, has_a_handle, a_tool, a_weapon*
GUN	gun, rifle, bazooka, shotgun	in top1	used_for_killing, a_weapon, made_of_metal, is_dangerous

Table 4: Qualitative analysis of predicted vectors (obtained by mapping from DS3 to FS1) for 10 concepts in the test set. Features annotated with an asterix(*) are not listed in the gold standard feature vector for the given concepts.

We compute the MAP score as follows: for each concept in the test set, we rank the features from the predicted feature vector in terms of their values, and measure the quality of this ranking with IR-based average precision, using the gold-standard feature set as the "relevant" feature set. The MAP score is then obtained by taking the mean average precision over the entire test set. Overall, the model seems to rank gold features highly, but the MAP score is certainly affected by the features which have not been seen in training (these account for 18.8% of the total number of features), because these will have a zero weight assigned to them, and so will be found at the end of the ranked feature list for that concept.

A qualitative evaluation of the top neighbours for predicted featural vectors 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. 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; or that jeans are worn on the lower body, as opposed to shawls which are worn on the shoulders. It is important to note that a production frequency of zero for a concept-feature pair in the McRae dataset does not necessarily mean that the feature is not a plausible property of the concept, but only that it is not one of the most salient features, since it was not produced by any of the human participants (*e.g.* the feature `has_teeth` has not been listed as a property of CAT in the McRae dataset, but it is clearly a plausible property of the CAT concept). Many of the top-predicted features for the concepts in the test set are plausible, even if they are not listed in the gold data (*e.g* `lives_in_water` for SEAWEED). This is yet another indication that the concept-feature pairs listed in the McRae dataset are not complete, meaning that there are salient features that apply to some concepts which have not been spelled out by the participants.

The ability to generalise feature representations to unseen concepts also means that these can now be evaluated on standard NLP tasks since we can obtain full coverage on the evaluation datasets. In order to show that the quality of the predicted vectors is in line with the state of the art on modelling concept similarity and relatedness, we computed the correlation on a subset of 1288 noun-noun pairs (485 words) from the MEN dataset (Bruni et al., 2014), leaving it to future work to test such transformations on different parts of speech like verbs or adjectives. It is important to mention that in the construction of this subset we also excluded all McRae concepts from MEN, because we didn't want any of that training data to occur in the test set. The mapping function was trained on all the concepts in the McRae dataset and then used to predict featural vectors for words in the MEN subset described above. A qualitative analysis of the predicted vectors show that they contain highly plausible features for words that are highly perceptual (*e.g.* the top predicted features for COOKIE are `is_round`, `is_edible`, `tastes_good`, `eaten_by_baking`), as opposed to words that are more abstract or don't rely on perceptual information (*e.g.* the top predicted features for LOVE are `an_animal`, `made_of_metal`, `is_sharp`). We obtain the best Spearman correlation (0.71) for the predicted featural vectors by training the mapping on the Mikolov vectors (DS5), the Spearman correlation of these vectors on the MEN subset being 0.75. The high correlation with the MEN scores shows that the featural vectors capture lexical similarity well, but suggest that rather than using them in isolation to construct a semantic model, they would be most helpful as an added modality in a multimodal semantic model.

4 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 psychological studies 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 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 are particularly predictive of which feature norms in feature space.

Acknowledgments

LF is supported by an EPSRC Doctoral Training Grant. EMV is supported by ERC Starting Grant DisCoTex (306920). SC is supported by EPSRC grant EP/I037512/1 and ERC Starting Grant DisCoTex (306920). We thank Douwe Kiela and the anonymous reviewers for their helpful comments.

References

Andrews, M., G. Vigliocco, and D. Vinson (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological review 116*(3), 463.

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

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

Barsalou, L. W., W. Kyle Simmons, A. K. Barbey, and C. D. Wilson (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences 7*(2), 84–91.

Bengio, Y., H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain (2006). Neural probabilistic language models. In *Innovations in Machine Learning*, pp. 137–186. Springer.

Bruni, E., N.-K. Tran, and M. Baroni (2014). Multimodal distributional semantics. *Journal Artificial Intelligence Research (JAIR) 49*, 1–47.

Clark, S. (2015). Vector space models of lexical meaning. *Handbook of Contemporary Semantics–second edition. Wiley-Blackwell*.

Devereux, B. J., L. K. Tyler, J. Geertzen, and B. Randall (2013). The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 1–9.

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

Hill, F., R. Reichart, and A. Korhonen (2014). Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics 2*, 285–296.

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

Kelly, C., B. Devereux, and A. Korhonen (2014). Automatic extraction of property norm-like data from large text corpora. *Cognitive Science 38*(4), 638–682.

Lazaridou, A., E. Bruni, and M. Baroni (2014, June). 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)*, Baltimore, Maryland, pp. 1403–1414. Association for Computational Linguistics.

McRae, K., G. S. Cree, M. S. Seidenberg, and C. McNorgan (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods 37*(4), 547–559.

Mevik, B.-H. and R. Wehrens (2007). The pls package: principal component and partial least squares regression in r. *Journal of Statistical Software 18*(2), 1–24.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Q. V. Le, and I. Sutskever (2013). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., W.-t. Yih, and G. Zweig (2013). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pp. 746–751. Citeseer.

Polajnar, T. and S. Clark (2014, April). 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*, Gothenburg, Sweden, pp. 230–238. Association for Computational Linguistics.

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

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

Sahlgren, M. (2006). *The Word-Space Model*. Dissertation, Stockholm University.

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

Vinson, D. P. and G. Vigliocco (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods 40*(1), 183–190.