Affordances in Grounded Language Learning

Stephen McGregor **KyungTae Lim**
LATTICE - CNRS & École normale supérieure / PSL
Université Sorbonne nouvelle Paris 3 / USPC
1, rue Maurice Arnoux, 92120 Montrouge, France
semcgregor@hotmail.com kyungtae.lim@ens.fr

Abstract

We present a novel methodology involving mappings between different modes of semantic representations. We propose distributional semantic models as a mechanism for representing the kind of world knowledge inherent in the system of abstract symbols characteristic of a sophisticated community of language users. Then, motivated by insight from ecological psychology, we describe a model approximating affordances, by which we mean a language learner's direct perception of opportunities for action in an environment. We present a preliminary experiment involving mapping between these two representational modalities, and propose that our methodology can become the basis for a cognitively inspired model of grounded language learning.

1 Introduction

Computational approaches to grounded language learning have typically involved mapping from perceptual to linguistic modalities through the application of complex information processing operations. Yu and Siskind (2013), for instance, use hidden Markov models to translate from *object tracks* to natural language descriptions of event observed in video clips. Likewise the ImageNet database has provided a platform for the productive application of deep neural network architectures for mapping between images and natural language labels (Krizhevsky et al., 2012). Significantly with regard to the ideas outlined here, Oh et al. (2017) describe a methodology for training an agent to construct novel sequences of actions based on analogies with previously learned strategies; the mechanism for learning a vocabulary of basic actions consists of a combination of convolutional and LSTM layers within a neural network.

Work of this nature highlights the state of the art in modelling technologies, and as an information engineering approach to meaningful tasks such as question answering and image labelling a significant contribution is made. This is arguably done, however, at the expense of presenting interpretable or indeed plausible models of the way that environmentally embedded agents use relatively scant exposure to a language speaking community in order to develop a lexicon that is rich and productive. In this regard, the conventional computational stance on grounded language learning embraces a view of the relationship between language and the world as a *symbol grounding problem*, by which abstract symbols susceptible to formal operations are somehow associated with perceptions and propositions: the hard work is done by a complex and philosophically opaque process of transforming signals into symbols, with the sense that computation by way of deep nets in some sense stands in for an inscrutable mind-brain gestalt.

As an alternative to this approach, Rączaszek-Leonardi et al. (2018) propose a *symbol ungrounding problem*: by this account, language begins as a semiotic structure with the representational scheme of a nascent language learner iconically and indexically aligned to embodied and embedded experiences of the world. This alignment is understood in terms of Gibson's (1979) notion of *affordances*, which we take to mean the direct perception of opportunities for action in an environment. The connection of language to opportunities for taking action on objects (and indeed the perception of language itself as an affordance for communication) creates a framework for understanding how abstract symbols begin as grounded complexes of multi-modal interactions with a language teacher and then gradually emerge as con-

straints on the way that a cognitive agent behaves in an environment (Rączaszek-Leonardi, 2012).

The strength of thinking of perception in general and language in particular in terms of affordances is that this moves away from the problem of the computational load associated with the spontaneous construction of contextually productive representational structures. For Clark (1997), affordances play a role in the an *action-oriented* model of cognition revolving around light-weight, environmentally situated representations, while Chemero (2009) proposes affordances as a mechanism for resolving the issue of the *mental gymnastics* inherent in a computational cognitive model. These approaches, which seek to place mind in the context of environmental embodiment and embeddedness, prefigure recent attempts to introduce affordances as a component of a cognitively oriented theory of language in which words can be mapped to denotations oriented towards action on objects in situations, and utterances themselves become opportunities for communication (Rączaszek-Leonardi and Nomikou, 2015).

Despite these valuable theoretical contributions, affordances have proved resistant to empirical modelling, not least because it is difficult to come up with a tractable scheme for representing a cognitive feature that is specifically conceived as an antidote to representational approaches to cognition. Our present objective is to begin to map a way towards the computational simulation of the role of affordances in language acquisition through interaction with an established linguistic community. In order to do this, we'll extract both statistical and syntactic information from a large-scale corpus to model two different modes of semantic representation, one geared towards the kind of world-knowledge inherent in the evolution of language on the time-scale of a community of language users, the other designed to reflect the way that an agent might encounter language grounded in the affordances of denoted objects.

In as much as we will be combining established distributional semantic techniques with likewise established syntactic analysis, this work can be broadly positioned in the context of other recent models. Cheng and Kartsaklis (2015), for instance, compound co-occurrence and syntactical information in order to generate word-embeddings enhanced for compositional tasks. Vulić (2017) likewise uses information about dependency relationships to map word embeddings from multiple languages into a shared vector space, achieving impressive results on cross-lingual versions of word similarity tasks. An important caveat regarding our own research, however, is that we are using syntactical information as a kind of stand-in for a simulation of the way that an agent might encounter words aligned with events involving objects: in the end, we would actually like to see the methodology outlined here as groundwork towards a model of language acquisition which specifically does not fall back on the kind of rich linguistic knowledge inherent in either vector space models or dependency parsers.

2 Modelling Affordances

On the one hand, as a model for the type of lexical semantic representation imbued with the productive world knowledge of an experienced language user, we propose distributional semantic vector space models (Clark, 2015), in which words are represented as points in high dimensional spaces where properties such as proximity and direction can relate to semantic phenomena such as relatedness and intension. On the other hand, as a model of the perception of objects mapped to linguistic units and at the same time perceived in terms of their potential for being acted upon, we suggest a rudimentary framework for associating denotations with events and related objects.

Distributional semantic word-vectors have the advantage of incorporating both world knowledge (Mikolov et al., 2013; Pennington et al., 2014) and at least the potential for compositionality (Mitchell and Lapata, 2010; Coecke et al., 2011) into a computationally tractable structure. They can also, importantly, be extrapolated in an essentially unsupervised way from large-scale textual data, allowing for the construction of an open ended lexicon characterised by the representation of semantic properties through geometric features. In terms of our framework for modelling a language-learning agent, we propose that this type of representation can stand in for interaction between the agent and a tutor acting as a conduit to the knowledge embedded in a language as developed by a community of language users (Smith et al., 2003).

In terms of actually instantiating this type of model, we apply the `word2vec` technique to learn a space of word embeddings (Mikolov et al.,

2013), applied across iterative observations of a cleaned-up version of Wikipedia.[1] We detect sentence boundaries, remove punctuation, render all characters lower-case, and, ignoring sentences of less than five words in length, apply the *skip-gram* methodology for learning to predict a 5-by-5 window of co-occurrence words around a given target word. The cleaned version of our corpus contains about 9.08 million word types representing roughly 1.87 billion word tokens spread across 87.2 million sentences.

As a model of the way that language is encountered in the environment by a novice language learner, we propose a representational scheme for affordances designed to reflect the actions and interactions that might be associated with an agent's early encounters with new objects. For present purposes, we will once again turn to a corpus-based technique for building representations: we traverse the same rendition of Wikipedia, seeking instances where words in our vocabulary are used as direct objects. We parse each sentence in the corpus using the Spacy parser; in instances of multi-word phrases, we treat the head as a candidate target word. For word types tagged as direct objects, we build up counts of corresponding predicates and associated subjects and indirect objects and then calculate probability distributions over the word types observed in each of these roles, so that affordances can be represented as a matrix of probability distributions over word types for each of these three grammatical classes for every word in a target vocabulary:

$$p(X|w) = \left(\frac{|x_{w,1}|}{|X_w|}, \frac{|x_{w,2}|}{|X_w|} ... \right) \quad (1)$$

$$w = (p(P|w), p(S|w), p(I|w)) \quad (2)$$

Here, a distribution $p(X|w)$ represents the discrete probabilities of words $(x_1, x_2...)$ being observed in a dependency relationship X with word w, calculated for predicates, subjects, and indirect objects (P, S, I) respectively. We take these grammatical features to correspond, at least in a rough sense, to the kind of thing that can be done with the corresponding target object, the things that do these things, and the things that can be affected by actions involving this object.

With these probability spaces established, we can compute the top words in terms of probability of observation in a particular grammatical role across all vocabulary words up to some arbitrary count. So, for instance, an affordance matrix for the word *taxi* built with three-element probability distributions would look like this (where PRO and YEAR are generic representations for personal pronouns and years respectively):

predicate	subject	ind. object
take = 0.587	PRO = 0.809	YEAR = 0.385
drive = 0.279	*who* = 0.112	*airport* = 0.365
hail = 0.134	*von* = 0.079	*station* = 0.250

3 A Small Experiment

Beginning with the framework described above, we first examine the degree to which our representations capture properties associated with the denotations of some basic nouns. In order to establish a small-scale vocabulary of objects, we turn to the tables of words exemplifying types of objects described by Rosch (1975) in her seminal work on conceptual prototypes. We choose the five words that were reported as most prototypical of five conceptual categories, as determined by a survey of a large number of respondents. The categories are VEHICLES, CLOTHING, TOOLS, FURNITURE, and FRUIT. Our objective for this preliminary work will be to establish representations of these object types in both a distributional semantic vector space and a probabilistic affordance space.

In order to explore the effectiveness of the conceptual spaces generated by the representational techniques described above, we first extract the word-vectors corresponding to our vocabulary from the `word2vec` distributional semantic model and perform k-means clustering on these, specifying a total of five target classes.[2] Results are reported in the WORD-VECTORS column in Table 1. While these clusters do not correspond exactly with human judgement, they do align somewhat with the expected delineations between object classes. The large cluster containing a mix of furniture and tools is characterised by words like *saw* and *ruler* which are presumably affected by a high degree of word sense ambiguity.

Next we explore the space of affordances. The representations in this space are, as described above, construed as matrices of probabilities. Specifically, we take the top 20 most likely words

[1] Implemented using the Gensim library for Python.

[2] Clustering is implemented using the `KMeans` algorithm from Python's sklearn library.

WORD-VECTORS	AFFORDANCES
automobile, truck, car, bus, taxi	**car**, *automobile, truck, taxi*
pants, shirt, dress, skirt, blouse	**dress**, *pants, shirt, skirt, blouse*
hammer, screwdriver	**hammer**, *table, bus, screwdriver, drill*
chair, sofa, couch, table, dresser, saw, ruler, drill	**chair**
orange, apple, banana, peach, pear	**orange**, *sofa, couch, dresser, apple, banana, peach, pear, saw, ruler*

Table 1: Clusters of word representations in distributional semantic and probabilistic affordance spaces. Word-vectors are clustered based on k-means clustering, and affordance representations are clustered based on a k-medoid algorithm, with the most cost-effective medoids indicated in bold.

for each grammatical class and generate probabilities for each word in each of these classes for each of our 25 object-words. In order to calculate the distance between two affordance representations, we take the Hellinger distance between two aligned probability distributions. This operation, which we take as a good quantification of the relationship between two distributions, results in a matrix of three dimensional vectors, each element corresponding to a grammatical class. So, for two vocabulary words a and b and a grammatical class c, the element of a vector representing the relationship between those two words can be described as follows, where h is the label for one of the top 20 words occurring in that grammatical class:

$$M_c(a,b) = \frac{1}{\sqrt{2}} \times \sqrt{\sum_{h=1}^{20} \left(\sqrt{p(a_h)} - \sqrt{p(b_h)} \right)^2} \tag{3}$$

We treat the set of three values corresponding to each target-to-target relationship as a distance vector, and so consider the distance between those two words to be simply the norm of that vector. With a distance matrix thus established, we use a k-medoids algorithm of our own design to cluster the affordance representations. We apply this measure because we are working from a matrix of distances, rather than from an explicit vector space; we might also consider, for instance, multi-dimensional scaling to project these representations into a vector space, but we consider the k-medoid approach to be appropriate for our present purposes. Results are reported in the right column of Table 1, with optimal medoids highlighted.

As with the clustering of word-vectors, the results here do not correspond perfectly with human judgements. We don't see this as necessarily being a problem, though: it would be strange, in fact, to expect a developing cognitive agent to categorically classify each object based on affordance-oriented interactions with an environment. So, for instance, fruits are compounded with some furniture and some tools in a single category orbiting the highly ambiguous term *orange*.

The crucial question is how we can effectively map between word-vectors, which we take to represent a kind of encyclopaedic knowledge of the world, and the affordances which are proposed as at least a rough model of the way that words are encountered by an early language learner. In order to explore this issue further, we construct a rudimentary neural network, mapping the 200 elements of each of our word-vectors onto the sets of probability distributions corresponding to affordances by way of a single dense softmax layer. This operation is in effect quite similar to a multi-class logistic regression, except that here we are attempting to learn to approximate an actual probability distribution rather than to simply reward the assignment of the highest score to a particular class. Formally, we map from a word-vector $\overrightarrow{v_w}$ to a probability distribution $p(x_n|w)$ associated with a word x_n observed participating with vocabulary word w as a member of grammatical class X by learning a weight matrix M, expressed here in terms of dot products with each row $\overrightarrow{m_{x_k}}$ associated with members $(x_1 ... x_{|X|})$ of class X:

$$p(x_n|w) = \frac{e^{\overrightarrow{v_w}^T \cdot \overrightarrow{m_{x_n}}}}{\sum_{k=1}^{k=|X|} e^{\overrightarrow{v_w}^T \cdot \overrightarrow{m_{x_k}}}} \tag{4}$$

A separate weight matrix is learned for each of the three grammatical classes associate with the objects that we seek to model.

As a basic test of the generality of this network, we perform a five-fold cross-validation, holding one term from each class out of the network construction process for each fold. Table 2 reports accuracy rates for this experiment, where a word-

	TOTAL	VEHICLES	CLOTHING	TOOLS	FURNITURE	FRUIT
word	0.12	0.0	0.2	0.2	0.2	0.0
class	0.64	0.8	0.6	0.6	0.2	1.0

Table 2: Accuracy rates for mapping from distributional semantic word-vectors to affordance matrices, from a word to the same word and from a word to another word of the same class.

vector is considered to map to the point in the space of affordance matrices that is closest based on Hellinger as technique described above. This experiment is designed to test the ability of this simple model to map between two different modes of semantic representation, one based on a large-scale analysis of the way that words occur in the context of a complex, developed vocabulary, the other utilising syntax to simulate the small-scale encounter of words as mapping to opportunities for action on corresponding denotations.

While the network generally fails to make exact word-to-word mappings, it is notable that it does, more often than not, manage to map a word-vector to an affordance representation corresponding to another word of at least the correct class. We suggest that this indicates there is some basic categorical information in word-vector representations that can be aligned with data about the way that objects are predictably encountered in the world.

4 What Next?

The development of a mapping between encyclopaedic and empirical lexical semantic representations described here is, in the end, not particular remarkable. We have in effect mapped from one statistical interpretation of a corpus to another. There is a large space of parameters to toggle: the parameters of the `word2vec` methodology for generating word-vectors, the number and choice of grammatical classes for our affordance space, the actual selection of target vocabulary words, and the network architecture for mapping between representational frameworks are just some of the factors inviting further experimentation.

Moreover, the experiments we carry out involve a small set of words distributed over a likewise small set of classes. This is in contrast to some of the more ambitious approaches to multi-modal tasks such as image labelling that have recently emerged, which can involve thousands of labels (Frome et al., 2013). What we are aiming at, though, is not so much an approach to information engineering as a first step towards modelling the way grounded language learning might happen for an environmentally situated agent.

To this end, there is ample room to reconsider the way in which we model affordances in the first place. The corpus-based technique described here is amenable to a computational approach, but ultimately it will be important to develop a more situated methodology. To this end, experiments on human-robot interaction conducted by (Gross and Krenn, 2016) have illustrated the way in which factors such as gaze and gesture are crucial features of early-stage linguistic interactions, and we suggest that a mechanism for representing these elements of communication is an important consideration in modelling grounded language learning. Likewise with a focus on robotic applications, Spranger and Steels (2014) have explored the way that the *ontogentic ritualisation* inherent in the phenotype of a community of language users plays an important role in human language learning.

Returning to more traditionally computational tasks, we finally propose that an affordance based model can play a useful role in mapping between low-level input from, for instance, the visual domain and more abstract linguistic representations. What we have described here might be conceived as one wing of the type of autoencoder network that has been successful in tasks involving image processing (Krizhevsky et al., 2012) and machine translation (Hill et al., 2016). Rather than treating the encoding at the locus of these networks as an arbitrarily abstract semantic representation, we propose that an effective system might involve encoding to and decoding from affordance type representations. The next step in exploring this hypothesis will be to experiment with mapping from images to affordances. We have no illusions that this will be an easy task, but we do think that we have established sufficient groundwork for carrying ahead with this line of research.

Acknowledgements

This work was supported by the ERA-NET Atlantis project.

References

Anthony Chemero. 2009. *Radical Embodied Cognitive Science*. The MIT Press, Cambridge, MA.

Jianpeng Cheng and Dimitri Kartsaklis. 2015. Syntax-aware multi-sense word embeddings for deep compositional models of meaning. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1531–1542. Association for Computational Linguistics.

Andy Clark. 1997. *Being There: Putting Brain, Body, and World Together Again*. MIT Press, Cambridge, MA.

Stephen Clark. 2015. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, pages 493–522. Wiley-Blackwell.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2011. Mathematical foundations for a compositional distributed model of meaning. *Linguistic Analysis*, 36(1-4):345–384.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129.

James J. Gibson. 1979. *The Ecological Approach to Visual Perception*. Houghton Miffline, Boston.

Stephanie Gross and Brigitte Krenn. 2016. The ofai multi-modal task description corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377. Association for Computational Linguistics.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, pages 1097–1105.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR Workshop*.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1439.

Junhyuk Oh, Satinder P. Singh, Honglak Lee, and Pushmeet Kohli. 2017. Zero-shot task generalization with multi-task deep reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2661–2670.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*.

Joanna Rączaszek-Leonardi. 2012. Language as a system of replicable constraints. In Howar Hunt Pattee and Joanna Rączaszek-Leonardi, editors, *Laws, Lanuage and Life*, pages 295–333. Springer.

Joanna Rączaszek-Leonardi and Iris Nomikou. 2015. Beyond mechanistic interaction: Value-based constraints on meaning in language. *Frontiers in Psychology*, 6(1579).

Joanna Rączaszek-Leonardi, Iris Nomikou, Katharina J. Rohlfing, and Terrence W. Deacon. 2018. Language development from an ecological perspective: Ecologically valid ways to abstract symbols. *Ecological Psychology*, 30(1):39–73.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104:192–233.

Kenny Smith, Henry Brighton, and Simon Kirby. 2003. Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in Complex Systems*, 6(4):537–558.

Michael Spranger and Luc Steels. 2014. Discovering communication through ontogenetic ritualisation. In *4th International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EPIROB)*, pages 14–19.

Ivan Vulić. 2017. Cross-lingual syntactically informed distributed word representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 408–414.

Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 53–63.