Affordance Extraction and Inference based on Semantic Role Labeling

Daniel Loureiro, Alípio Mário Jorge
LIAAD - INESC TEC
Faculty of Sciences - University of Porto, Portugal
dloureiro@fc.up.pt, amjorge@fc.up.pt

Abstract

Common-sense reasoning is becoming increasingly important for the advancement of Natural Language Processing. While word embeddings have been very successful, they cannot explain which aspects of 'coffee' and 'tea' make them similar, or how they could be related to 'shop'. In this paper, we propose an explicit word representation that builds upon the Distributional Hypothesis to represent meaning from semantic roles, and allow inference of relations from their meshing, as supported by the affordance-based Indexical Hypothesis. We find that our model improves the state-of-the-art on unsupervised word similarity tasks while allowing for direct inference of new relations from the same vector space.

1 Introduction

The word representations used more recently in Natural Language Processing (NLP) have been based on the Distributional Hypothesis (DH) (Harris, 1954) — "words that occur in the same contexts tend to have similar meanings". This simple idea has led to the development of powerful word embedding models, starting with Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) and later, the popular word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) models. Although, effective at quantifying the similarity between words (and phrases) such as 'tea' and 'coffee', they cannot relate that judgement to the fact that both can be sold, for instance. Furthermore, current representations can't inform us about possible relations between words occurring in mostly distinct contexts, such as using a 'newspaper' to cover a 'face'. While there have been substantial improvements to word embedding models over the years, these shortcomings have endured (Camacho-Collados and Pilehvar, 2018).

Word Pairs	**Affordances**
(w_1, w_2)	**(w_1 as ARG0, w_2 as ARG1)**
shop, tea	sell, import, cure
doctor, patient	diagnose, prescribe, treat
newspaper, face	cover, expose, poke
man, cup	drink, pour, spill

Table 1: Results from affordance meshing (coordination) using automatically labelled semantic roles.

Glenberg et al. (2000) identified these issues soon after LSA was introduced, and cautioned that high-dimensional word representations, such as those based on the DH, lack the necessary grounding to be proper semantic analogues. Instead, Glenberg proposed the Indexical Hypothesis (IH) which supports that meaning is constructed by (a) indexing words and phrases to real objects or perceptual, analog symbols; (b) deriving affordances from the objects and symbols; and (c) meshing the affordances under the guidance of syntax. Following Glenberg et al. (2000), this work considers an object's affordances as its possibilities for action constrained by its context, including actions which may not be directly perceived, which differs slightly from Gibson (1979)'s original definition. Even though the language grounding advocated by the IH is beyond the reach of NLP by itself, we believe that its representation of meaning through affordances can still be captured to a useful extent.

Our contribution[1] is a word-level representation that allows for the affordance correspondence and meshing supported by the IH. These affordances are approximated from occurrences of semantic roles in corpora through an adaptation of models based on the DH. Our work is motivated by two observations: (1) a pressing need to integrate common-sense knowledge in NLP mod-

[1] Code, data and demo: https://a2avecs.github.io

Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 91–96
Brussels, Belgium, November 1, 2018. ©2018 Association for Computational Linguistics

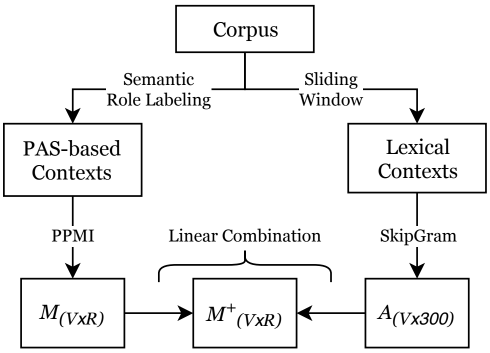

Figure 1: Outline of model pipeline.

els and (2) recent improvements to Semantic Role Labeling (SRL) have made affordance extraction from raw corpora sufficiently reliable. We find that our model (A2Avecs) performs competitively on word similarity tasks while enabling novel 'who-does-what-to-whom' style inferences (Table 1).

2 Related Work

This work is closely related to the research area of selectional preferences, where the goal is to predict the likelihood of a verb taking a certain argument for a particular role (e.g. likelihood of *man* being an *agent* of *drive*). Most notably, Erk et al. (2010) proposed a distributional model of selectional preferences that used SRL annotations as a primary set of verb-role-arguments from which to generalize using word representations based on the DH and several word similarity metrics. Progress in selectional preferences is usually measured through correlations with human thematic fit judgements and, more recently, neural approaches (de Cruys, 2014; Tilk et al., 2016) obtained state-of-the-art results.

While this work shares some of these same elements (i.e. SRL and word embeddings), they are used to predict potential affordances instead of selectional preferences. Consequently, our representations are designed to enable the meshing proposed by the IH, allowing us to infer affordances that would not be likely under a selectional preference learning scheme (e.g. *newspaper-cover-face* from Table 1). Additionally, this work is concerned with showing that information derived from SRL is complementary to information derived from DH methods, and thus focuses its evaluation on tasks related to lexical similarity rather than thematic fit correlations.

3 Method

Our word representations are modelled using Predicate-Argument Structures (PASs). These structures are obtained through SRL of raw corpora, and used to populate a sparse word/context co-occurrence matrix W where roles serve as contexts (features), and argument spans serve as the co-occurrence windows. The roles are predicates specified by argument type (e.g. eat|ARG0) and used in place of affordances. See Table 2 for a comparison of this context definition with the traditional lexical definition.

	Context	**Words**
Role	drinks|ARG0	John
	drinks|ARG1	red, wine
	drinks|ARGM-MNR	slowly
Adjacency	John	drinks, red
	drinks	John, red, wine
	red	John, drinks, wine, slowly
	wine	drinks, red, slowly
	slowly	red, wine

Table 2: Different context definitions applied to the sentence 'John drinks red wine slowly'. Top: Our proposed definition; Bottom: Lexical adjacency definition (with window size of 2).

After computing our co-occurrence matrix we follow-up with the additional steps employed by traditional bag-of-words models. We use Positive Pointwise Mutual Information (PPMI) to improve co-occurrence statistics, as used successfully by Bullinaria and Levy (2007); Levy and Goldberg (2014b), and maintain explicit high-dimensional representations in order to preserve the context information required for affordance meshing. Previous works, such as Levy and Goldberg (2014a) and Stanovsky et al. (2015), have also produced word representations from syntactic context definitions (dependency parse trees and open information extraction, respectively) but have opted for following-up with the word2vec's SkipGram (SG) model, presumably influenced by a much higher number of contexts in their approaches.

We reduce the sparsity of our explicit PPMI matrix by linear combination and interpolation of semantically related vectors. The semantic relatedness is obtained from the cosine similarity of SG vectors. As evidenced by Baroni et al. (2014), SG seem best suited for estimating relatedness (or association). These steps are further described in remainder of this section (See Fig. 1).

3.1 Extracting PASs

We use the AllenNLP (Gardner et al., 2017) implementation of He et al. (2017) state-of-the art SRL to extract PASs from an English Wikipedia dump from April 2010 (1B words). Since the automatic identification of predicates by an end-to-end SRL may produce erroneous results, we ensure that these predicates are valid by restricting them to the set of verbs tagged on the Brown corpus (Francis and Kucera, 1979). We also use the spaCy parser (Honnibal and Montani, 2017) to reduce each argument phrase to its head noun phrase, reducing the dilution of the more relevant noun and predicate co-occurrence statistics (See Fig. 2). Additionally, we lemmatize the predicates (verbs) to their root form using WordNet's Morphy Lemmatizer (Miller, 1992). Finally, we trim the vocabulary size and the number of roles by discarding those which occur less than 100 times, and consider only core and adjunct argument types. The result is a set C of observed contexts, such as <chase|ARG1, (the, cat)>, used to populate W.

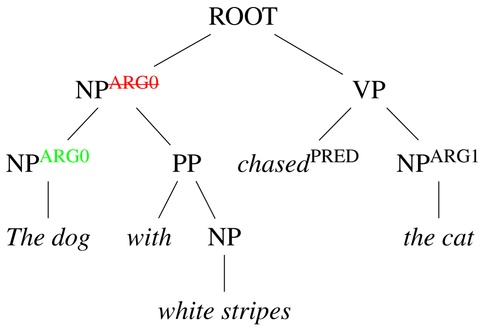

Figure 2: Parse tree for the sentence 'The dog with white stripes chased the cat.'. The label for ARG0 is repositioned to the smaller subtree.

3.2 Argument-specific PPMI

The authors of PropBank (Kingsbury and Palmer, 2002), which provides the annotations used for learning SRL, state that arguments are predicate-specific. Still, they also acknowledge that there are some trends in the argument assignments. For instance, the most consistent trends are that ARG0 is usually the agent, and ARG1 is the direct object or theme of the relation. This realisation leads us to adapting the PPMI measure to better account for the correlations between roles of the same argument types. Thus, we segment C by argument type, and apply PPMI independently considering

PST	PMI	THR	RND	MEN	SP
L				.249 (1)	98.71
L+H				.309 (47)	99.41
L+H	P			.606 (47)	99.58
L+H	A+P			.611 (47)	99.59
L+H	A+P	0.5		N/A[a]	< 41.2
L+H	A+P	0.5	HDR	**.687 (0)**	98.21
L+H	A+P	0.6	HDR	.654 (14)	97.98
L+H	A+P	0.4	HDR	.668 (0)	98.77

[a] Failed after using too much memory.

Table 3: Sensitivity/Impact analysis for some parameters of our approach.
Legend: **PST**: Post-Processing (L: Lemmatization; H: Head noun phrase isolation); **PMI**: PMI Variations (P: PPMI; A: Arg-specific PPMI); **THR**: Similarity threshold (tested values); **RND**: Rounding (HDR: Half down rounding); **MEN**: MEN-3K task (Spearman correlation, #OOV failures); **SP**: Sparsity (percentage of zero values on a 155Kx18K matrix).

each C_{ARG}, such that for each $W_{w,p}$:

$$PPMI(w,r) = max(PMI(w,r),0))$$

$$PMI(w,r) = log\frac{f(w,r)}{f(w)f(r)} = log\frac{\#(w,r)\cdot|C_{ARG}|}{\#(w)\cdot\#(r)}$$

where w is a word from the vocabulary V, r is a role (context) from the set R of the same argument type as C_{ARG}, and f is the probability function. The resulting matrix $M = PPMI(W)$ maintains the dimensions W and is slightly sparser.

3.3 Leveraging Association

The constraints imposed by SRL yield a very reduced number of PAS-based contexts that can be extracted from a corpus, in comparison to lexical adjacency-based contexts. Moreover, the post-processing steps we perform, while otherwise beneficial (see Table 3), further trim this information. To mitigate this issue, we also compute an embedding matrix A (see Section 3 for parameters), using the state-of-the-art lexical-based SG model of Bojanowski et al. (2017), and use those embeddings to obtain similarity values that can be used to interpolate missing values in M, through weighted linear combination. This way, existing vectors are re-computed as:

$$\vec{v}_w = \frac{\vec{v}_1 * \alpha_1 + ... + \vec{v}_n * \alpha_n}{\sum_{i=1}^{n}\alpha_i}$$

with α_i defined as:

$$\alpha_i = \begin{cases} \frac{A_{\vec{v}_w} \cdot A_{\vec{v}_i}}{\|A_{\vec{v}_w}\| \|A_{\vec{v}_i}\|} = cos_A(\vec{v}_w, \vec{v}_i), \\ \qquad\qquad \text{if } cos_A(\vec{v}_w, \vec{v}_i) > 0.5 \\ 0, \quad \text{otherwise} \end{cases}$$

where cos_A corresponds to the cosine similarity in the SG representations.

The similarity threshold is tested on a few natural choices (0.5 ± 0.1) and validated from results on a single word similarity task (see Table 3). This approach is also used to define representations for words that are out-of-vocabulary (OOV) for M, but can be interpolated from related representations, similarly to Zhao et al. (2015). In conjunction with the interpolation, we apply half down rounding to the vectors, before and after re-computing them, so that our representations remain efficiently sparse while benefitting from improved performance. Finally, we apply a quadratic transformation to enlarge the influence of meaningful co-occurrences, obtaining $M^+ = interpolate(M, A)^2$.

3.4 Inferring Relations

The examples shown in Table 1 are easily obtained with our model through a simple procedure (see Algorithm 1) that matches different arguments of the same predicates. As was the case with Arg-specific PPMI, this procedure is made possible by the fact that a significant portion of argument assignments remain consistent across predicates.

Algorithm 1 Affordance Meshing Algorithm

```
procedure INFERENCE($M^+, w_1, w_2, a_1, a_2$)
    $relations \leftarrow []$
    $\vec{v}_1, \vec{v}_2 \leftarrow get\_vec(w_1, M^+), get\_vec(w_2, M^+)$
    for $f_1 \in features(\vec{v}_1) \wedge arg(f_1) = a_1$ do
        for $f_2 \in features(\vec{v}_2) \wedge arg(f_2) = a_2$ do
            if $pred(f_1) = pred(f_2)$ then
                $relations.add((f_1 * f_2, pred(f_1)))$
    return $sorted(relations)$
```

4 Evaluation and Experiments

The A2Avecs model introduced in this paper is used to generate 155,183 word vectors of 18,179 affordance dimensions. This section compares our model with lexical-based models (word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and fastText (Bojanowski et al., 2017)) and other syntactic-based models (Deps (Levy and Goldberg, 2014a) and OpenIE (Stanovsky et al., 2015)). We're using Deps and OpenIE embeddings that the respective authors trained on a Wikipedia corpus and distributed online. Lexical models were trained using the same parameters, wherever applicable: Wikipedia corpus from April 2010 (same as mentioned in section 2.1); minimum word frequency of 100; window size of 2; 300 vector dimensions; 15 negative samples; and ngrams sized from 3 to 6 characters.

We also show that our approach can make use of high-quality pretrained embeddings. We experiment with a fastText model pretrained on 600B tokens, referred to as 'fastText 600B' in contrast with the fastText model trained on Wikipedia.

4.1 Model Introspection

The explicit nature of the representations produced by our model makes them directly interpretable, similarly to other sparse representations such as Faruqui and Dyer (2015b). The examples presented at Table 4 demonstrate the relational capacity of our model, beyond associating meaningful predicates. In this introspection we highlight the top role contexts for a set of words, inspired by (Levy and Goldberg, 2014a) which presented the top syntactic context for the same words, and note that this introspection produces results that should correspond to Erk et al. (2010)'s inverse selectional preferences.

Our online demonstration provides access to additional introspection queries, such as top words for given affordances, or which affordances are most distinguishable between a pair of words (determined by absolute difference).

batman	**hogwarts**	**turing**
foil|ARG0	ambush|ARGM-MNR	travel|ARGM-TMP
flirt|ARGM-MNR	rock|ARGM-LOC	pass|ARGM-ADV
apprehend|ARG0	express|ARG0	solve|ARG0
subdue|ARG0	prevent|ARGM-LOC	simulate|ARG1
rescue|ARGM-DIR	expel|ARG2	prove|ARG1
florida	**object-oriented**	**dancing**
base|ARGM-MNR	define|ARG1	dance|ARG0
vacation|ARG1	define|ARG0	dance|ARGM-MNR
reside|ARGM-DIS	use|ARG1	dance|ARGM-LOC
fort|ARG1	implement|ARG1	dance|ARGM-ADV
vacation|ARGM-LOC	express|ARGM-MNR	dance|ARG1

Table 4: Words and their top role contexts. Using the same words from the introspection of (Levy and Goldberg, 2014a) to clarify the difference in the representations of both approaches.

Context	Model	SL-666	SL-999	WS-SIM	WS-ALL	MEN	RG-65
Lexical	word2vec	.426	.414	.762	.672	.721	.793
	GloVe	.333	.325	.637	.535	.636	.601
	fastText (A)	.426	.419	**.779**	**.702**	**.751**	.799
Syntactic	Deps	**.475**	**.446**	.758	.629	.606	.765
	Open IE	.397	.390	.746	.696	.281	.801
	A2Avecs (M^+)	.461	.412	.734	.577	.687	**.802**
	A2Avecs (SVD(M^+))	.436	.386	.672	.509	.599	.789

Lexical SOTA	fastText 600B (A)	.523	.504	.839	**.791**	**.836**	**.859**
Intp. w/SOTA	A2Avecs (M^+)	.513	.468	.780	.619	.744	.814
Intp. & Conc.	A2Avecs ($M^+ \parallel A$)	**.540**	**.521**	**.846**	.771	.829	.857
Deps Conc.	Deps $\parallel A$	.524	.503	.818	.752	.770	.835

Table 5: Spearman correlations for word similarity tasks (see Faruqui and Dyer (2014) for task descriptions). Top section shows results from training on the Wikipedia corpus exclusively. Bottom section shows results where we used SG embeddings (A) trained on a larger corpus for performing interpolation and concatenation on the same set of roles used above. For comparison, we also show results for Deps concatenated with those embeddings.

4.2 Word Similarity

The results presented on Table 5 show that our model can outperform lexical and syntactic models, in spite of maintaining an explicit representation. In fact, applying Singular Value Decomposition (SVD) to obtain dense 300-dimensional embeddings degrades performance. We achieve best results with the concatenation of the fastText 600B vectors with our model interpolated using those same vectors for the vocabulary $V_{M^+} \cap V_A$, after normalizing both to unit length (L_2). Interestingly, the same concatenation process with Deps embeddings doesn't seem as beneficial, suggesting that our representations are more complementary.

5 Conclusion

Our results suggest that semantic similarity can be captured in a vector space that also allows for the inference of new relations through affordance-based representations, which opens up exciting possibilities for the field. In the process, we presented more evidence to support that information obtained from SRL is complementary to information obtained from adjacency-based contexts, or even contexts based on syntactical dependencies. We believe this work helps bridge the gap between selectional preferences and semantic plausibility, beyond frequentist generalizations based on the DH. In the near term, we expect that specific tasks such as Entity Disambiguation and Coreference can benefit from these representations. With further developments, semantic plausibility assessments should also be useful for more broad tasks such as Fact Verification and Story Understanding.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior research methods*, 39 3:510–26.

José Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning.

Tim Van de Cruys. 2014. A neural network approach to selectional preference acquisition. In *EMNLP*.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36:723–763.

Manaal Faruqui and Chris Dyer. 2014. Community evaluation and exchange of word vectors at wordvectors.org. In *ACL*.

Manaal Faruqui and Chris Dyer. 2015b. Non-distributional word vector representations. In *ACL*.

W. Nelson Francis and Henry Kucera. 1979. The Brown Corpus: A Standard Corpus of Present-Day Edited American English. Brown University Liguistics Department.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

J. J. Gibson. 1979. The ecological approach to visual perception. *Brain and language*.

Arthur M. Glenberg, David A. Robertson, Brianna Benjamin, Jennifer Dolland, Jeanette Hegyi, Katherine V Kortenkamp, Erik Kraft, Nathan Pruitt, Dana Scherr, and Sara Steinberg. 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning.

Zellig Harris. 1954. Distributional structure. page 10(23):146162.

Luheng He, Kenton Lee, Mike Lewis, and Luke S. Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *ACL*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *LREC*.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *ACL*.

Omer Levy and Yoav Goldberg. 2014b. Linguistic regularities in sparse and explicit word representations. In *CoNLL*.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*.

George A. Miller. 1992. Wordnet: A lexical database for english. *Commun. ACM*, 38:39–41.

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

Gabriel Stanovsky, Ido Dagan, and Mausam. 2015. Open ie as an intermediate structure for semantic tasks. In *ACL*.

Ottokar Tilk, Vera Demberg, Asad B. Sayeed, Dietrich Klakow, and Stefan Thater. 2016. Event participant modelling with neural networks. In *EMNLP*.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning translation models from monolingual continuous representations. In *HLT-NAACL*.