A Distributional Semantic Model of Visually Indirect Grounding for Abstract Words

Akira Utsumi
Department of Informatics / Artificial Intelligence eXploration Research Center
The University of Electro-Communications, Tokyo 182-8585, Japan
utsumi@uec.ac.jp

Abstract

Recent studies on abstract words have proposed an indirect grounding view that abstract words are grounded indirectly through linguistic interdependency among amodal and modal symbols. In this paper, we model the indirect grounding view by means of multimodal distributional semantics. In the proposed model, an abstract word is represented by both its textual vector and the visual vectors of its semantically related concrete words. A simulation experiment demonstrated that the indirect grounding model achieved better performance in predicting human conceptual representation of abstract words. This finding indicates the plausibility of indirect grounding as a cognitive mechanism of representing abstract words.

1 Introduction

Since Harnad (1990) pointed out the symbol grounding problem, cognitive science research on embodied cognition has demonstrated that the meaning or conceptual representation of words is largely grounded in perceptual or sensorimotor experiences. However, abstract words pose a serious challenge to the embodied theory of language, because it is difficult to see how representations grounded in perceptual experience can capture the content of abstract words such as *truth* and *justice* (Dove, 2015). In recent years, this challenge has become particularly topical and some special issues on abstract concepts have been published (Bolognesi & Steen, 2018; Borghi et al., 2018).

One of the important questions to be addressed by research on abstract concepts is how language and perceptual experience contribute to shaping our meaning representation of abstract words. It has been accepted that language is much more important for representing abstract concepts (Borghi et al., 2017; Dove, 2018), while the embodied theory of cognition claims that abstract concepts are also grounded in perception and action (Barsalou, 1999; Gibbs, 2006). Recent research trend in the studies on abstract concepts is the emergence of hybrid views that combine language and perceptual experience (Dove, 2018; Thill et al., 2014; Louwerse, 2011, 2018). For example, Thill et al. (2014) propose a "division of labor" approach between a perceptual layer that associates basic, concrete concepts with perceptual features and a relational layer that grounds more complex and abstract concepts in relation to basic concepts. Louwerse (2011, 2018) also proposes the symbol interdependency hypothesis, according to which language comprehension is symbolic through interdependencies of amodal linguistic symbols, while it is indirectly embodied through the references linguistic symbols make to perceptual representations. The basic idea underlying these "indirect grounding" views is that abstract words are grounded in sensorimotor or perceptual experiences, but the grounding is indirect, rather than direct in the case of many concrete words.

In this paper, we examine the validity and potential ability of the indirect grounding view of abstract words by means of multimodal distributional semantics (Bruni et al., 2014; Silberer et al., 2017). In multimodal distributional semantics, textual information is integrated with perceptual information computed directly from nonlinguistic inputs such as visual (Bruni et al., 2014; Kiela et al., 2014) and auditory (Kiela & Clark, 2015) ones. We focus on visual images as a source of perceptual information, and model a mechanism of indirect grounding via language. In the model, the visual vector

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

of an abstract word is computed from the visual images of concrete words semantically associated with the abstract word, rather than directly from the visual images tagged with the abstract word.

2 Distributional Semantic Model of Indirect Grounding

We assume that the vocabulary V is divided into concrete words V_C and abstract words V_A. Each word $w_i \in V$ has a textual vector $\vec{t}_i \in DSM_T$ trained from a text corpus and a visual vector $\vec{v}_i \in DSM_V$ computed from images for w_i. We build an indirect grounding model DSM_G in which a word is represented by a pair $(\vec{t}_i, \vec{g}_i)$ of a textual vector $\vec{t}_i$ and an indirectly grounded visual vectors $\vec{g}_i$ defined as follows (Takano & Utsumi, 2016):

$$\vec{g}_i = \begin{cases} \vec{v}_i & \text{(for a concrete word } w_i \in V_C\text{)} \\ \sum_{w_j \in SN(w_i)} \vec{v}_j / |SN(w_i)| & \text{(for an abstract word } w_i \in V_A\text{)} \end{cases} \quad (1)$$

where $SN(w_i) \subset V_C$ is a set of k semantically nearest neighbors of (i.e., k concrete words semantically related to) w_i. Semantic neighbors (or *mediator* words) of an abstract word w_i are determined by first selecting $K(> k)$ semantic neighbors of w_i from the whole vocabulary V and then selecting k nearest concrete words from the set of K neighbors. Semantic neighbors are computed using cosine similarity in the textual model DSM_T. The reason for limiting K neighbors before selecting k concrete words is that some highly abstract words (e.g., *truth*, *wisdom*) may not have semantically similar concrete words, and in this case it is more appropriate not to consider a visual representation.

3 Evaluation Experiment

To explore the cognitive plausibility of the indirect grounding model, we evaluated how accurately the model can predict human conceptual representation. In the evaluation experiment, the prediction function from multimodal word vectors $(\vec{t}_i, \vec{g}_i)$ to the conceptual representation $\vec{y}_i$ is trained using the feed-forward neural network shown in Figure 1, and the conceptual representation of untrained words is predicted by the trained network. To compare with our indirect grounding model, we also conducted the same experiment using a textual model DSM_T (i.e., the visual layers were not used), a visual model DSM_V (i.e., the textual layers were not used), and a standard multimodal (hybrid) model $DSM_H = \{(\vec{t}_i, \vec{v}_i) \mid \vec{t}_i \in DSM_T, \vec{v}_i \in DSM_V\}$.

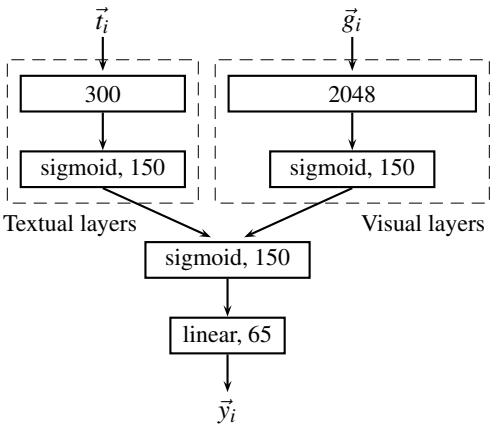


Figure 1: Neural network for evaluation.

Human conceptual representation. As a target human conceptual representation $\vec{y}_i$, we used Binder et al.'s (2016) brain-based semantic vectors of 535 words, comprising 434 nouns, 62 verbs and 39 adjectives. [1] This representation comprises 65 neurobiologically plausible attributes in 14 domains whose neural correlates have been well documented. Each word is represented as a 65-dimensional vector and each dimension corresponds to one of these attributes. Each value of the brain-based vectors represents the salience of the corresponding attribute, which is calculated as the mean salience rating on a 7-point scale ranging from 0 to 6.

Multimodal vector. Textual vectors $\vec{t}_i$ were trained on the Corpus of Contemporary American English (COCA), which included 0.56G word tokens. Words that occurred less than 30 times in the corpus were ignored, resulting in the training vocabulary of 108,230 words. As a training model for textual vectors, we used skip-gram with negative sampling (SGNS; Mikolov et al., 2013). We set the vector dimension $d = 300$ and the window size $w = 10$. The choice of corpus, training model and parameter values was determined considering the result of the similar experiment (Utsumi, 2018).

To compute visual vectors $\vec{v}_i$, we collected 20 images using *Flickr* image retrieval for each word. Each image was entered into the pretrained ResNet152-hybrid1365 model (Zhou et al., 2018) [2] and

[1] http://www.neuro.mcw.edu/semanticrepresentations.html

[2] https://github.com/CSAILVision/places365

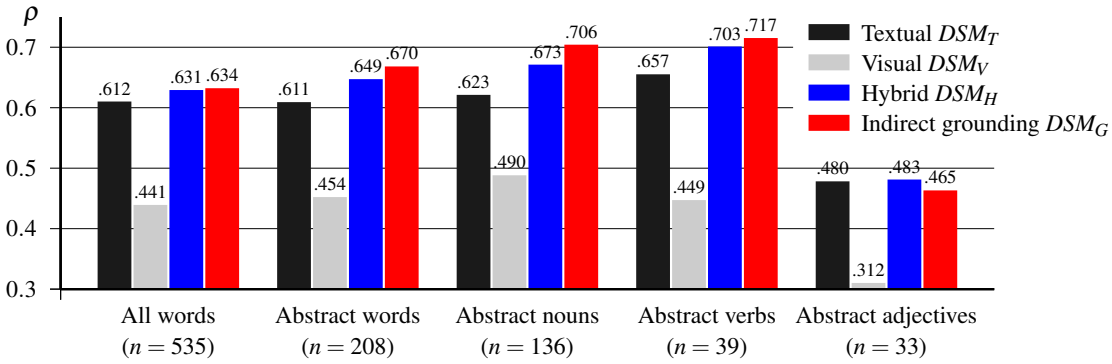


Figure 2: Mean correlation over words for the indirect grounding model and baseline models.

Table 1: Abstract words whose prediction performance was improved or impoverished by the indirect grounding model. The degree of improvement was computed as $\rho(DSM_G) - \max(\rho(DSM_T), \rho(DSM_H))$. **Left:** The best five and worst five abstract words and their concreteness ratings, correlations, and neighbor words used for computing indirect visual vectors. **Right:** The list of the best 20 and worst 10 abstract words.

		Correlation ρ			
Word	Conc.	DSM_T	DSM_H	DSM_G	Concrete mediator words
verb	2.85	.200	.218	.626	cryptogram, paragraph, comma
patent	3.32	.267	.240	.561	drugmaker, company, product
patient	2.50	.398	.424	.630	doctor, physician, hospital
matinee	3.78	.495	.607	.804	theater, concert, playhouse
diplomat	3.67	.599	.589	.793	ambassador, embassy, journalist
joviality	2.28	.686	.740	.460	stroking, smile, slouching
white	3.89	.391	.250	.111	red, yellow, colored
party	3.89	.590	.517	.255	democrat, pooper, dinner
fix	2.93	.348	.776	.425	screw, lock, broken
friendly	2.32	.637	.804	.409	handshake

Best 20 words:
verb, patent, patient, matinee, diplomat, evening, advantage, live, snub, moral, activist, soft, rumor, woe, zone, interview, grief, steal, happy, tax
Worst 10 words:
black, like, bribe, terrorist, end, joviality, white, party, fix, friendly

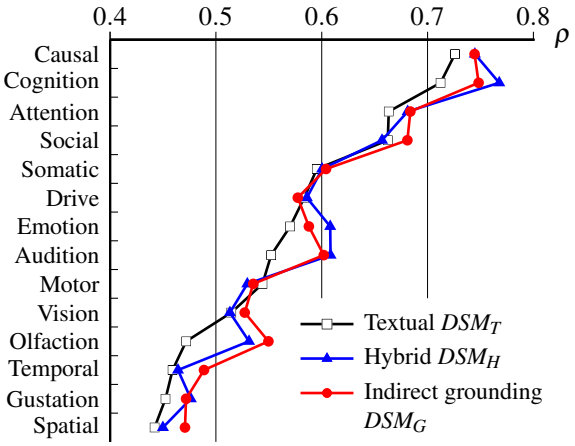
a 2048-dimensional feature vector was extracted from the last layer. Finally, the visual vector $\vec{v}_i$ was computed as the centroid of feature vectors of 20 images.

To determine concrete and abstract words, we used Brysbaert et al.'s (2014) concreteness ratings (on a 5-point scale ranging from 1 to 5) for 37,058 English words. Out of them, 27,856 words were chosen for a vocabulary for indirect grounding such that they were included in the training vocabulary and associated with at least 20 images. Each word was judged as abstract if its concreteness rating was less than a threshold θ_C. The threshold was estimated as $\theta_C = 3.9$ using grid search with a step size of 0.1, and as a result 208 out of 535 words in the vocabulary of Binder et al.'s (2016) representation were judged as abstract. The parameters K and k for computing $\vec{g}_i$ were also determined using grid search with $\theta_C = 3.9$, resulting in $K = 100$ and $k = 3$.

Training and prediction. Training and prediction were performed by a cross validation procedure. All 535 words were classified into 47 semantic categories provided by Binder et al. (2016). For each of the 47 categories, we trained the prediction function using all words in the remaining 46 categories and predicted brain-based vectors for words in that category. By repeating this procedure with every category as a target, we obtained $\widehat{B}$ with estimated brain-based vectors as rows.

Prediction performance of the estimated vectors was measured using Spearman's rank correlation ρ between the estimated brain-based matrix $\widehat{B}$ and the original matrix B. We performed two analyses: row-wise and column-wise matrix correlation. The row-wise correlation indicates the prediction accuracy for each word, while the column-wise correlation indicates the accuracy for each attribute.

Result. Figure 2 shows the result of mean word correlations. For mean correlations across all words, the indirect grounding model achieved higher performance than textual-only and visual-only models, but its performance did not differ from that of the hybrid model. For abstract words, however, the indirect grounding model achieved 3.2% improvement over the hybrid model and 9.7%



Attribute (Domain)	DSM_T	DSM_H	DSM_G
Landmark (Spatial)	.438	.438	.504
Dark (Vision)	.261	.329	.375
LowerLimb (Motor)	.450	.428	.492
Duration (Temporal)	.409	.394	.448
Fast (Vision)	.498	.515	.553
Communication (Social)	.655	.694	.726
Scene (Spatial)	.617	.643	.674
Color (Vision)	.568	.626	.654
Away (Spatial)	.372	.405	.432
Sound (Audition)	.657	.675	.695

Figure 3: **Left:** Mean correlation per attribute domain. **Right:** The list of the top 10 attributes and their correlations in terms of the degree of improvement $\rho(DSM_G) - \max(\rho(DSM_T), \rho(DSM_H))$.

improvement over the textual model. In particular, top five abstract words listed in Table 1 were improved by 187% $\sim$ 32% as compared to the hybrid model. These results indicate that indirect grounding is plausible as a cognitive model of human conceptual knowledge of abstract words.

Figure 2 also suggests that the effect of visually indirect grounding differs among word classes. Although the accuracy for abstract nouns and verbs was improved by the indirect grounding model, abstract adjectives were predicted less accurately than by the hybrid and textual models. One possible reason for this result lies in the nature of neighbor words of adjectives. As shown in the case *white* of Table 1, neighbor words of adjectives tend to be adjectives but the meaning of adjectives is difficult to train from images. It may worsen the precition accuracy for abstract adjectives.

Figure 3 shows mean column-wise correlations per attribute domain. This result can be interpreted as indicating what type of information is likely to be represented by the indirect grounding model. Overall, causal, cognitive and social attributes, which primarily characterize abstract words (Binder et al., 2016), were predicted more accurately than perceptual, motor, and spatiotemporal attributes. Both indirect and hybrid models yielded higher correlations for almost domains than the textual model. The indirect grounding model improved the performamnce for *Social*, *Vision*, *Olfaction*, *Temporal* and *Spatial* domains than both textual and hybrid models. The reason for better representing social and spatiotemporal information might be that ResNet152 model used for computing visual vectors was pretrained with not only ImageNet dataset for object recognition but also Places image dataset for scene recognition.

4 Discussion

In this paper, we have demonstrated that the distributional semantic model for indirect grounding improved the representation performance of abstract words. This finding lends support to the indirect grounding view that abstract words are indirectly grounded through linguistic interdependency.

However, the model presented in this paper is so simple that we must modify it in several ways to provide more conclusive evidence for the role of language in grounding abstract words. One important issue to address for further work is a method for choosing mediator words whose visual vectors form the grounded representation of abstract words. One effective method is to limit the vocabulary of concrete words from which mediator words are chosen. Highly concrete words or other kinds of word set such as "Minimal Grounding Set" (Vincent-Lamarre et al., 2016) can be used as a candidate set of mediator words. Limiting word class to noun may also be an effective method, in particular for abstract adjectives. To improve the quality of visual vectors, we can use image dispersion (Kiela et al., 2014), namely the degree of similarity among images for the same word. Furthermore, different methods for selecting neighbor words worth pursuing; some empirical findings have shown that affective or emotional experiences are crutial in processing abstract concepts (Kousta et al., 2011; Vigliocco et al., 2014) and thus affectively similar words are likely to be good mediators of indirect grounding.

Acknowledgments

This research was supported by JSPS KAKENHI Grant Numbers JP15H02713 and SCAT Research Grant.

References

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*(3–4), 130–174.

Bolognesi, M., & Steen, G. (2018). Abstract concepts: Structure, processing, and modeling. *Topics in Cognitive Science*, *10*(3), 490–500.

Borghi, A. M., Barca, L., Binkofski, F., & Tummolini, L. (2018). Varieties of abstract concepts: Development, use and representation in the brain. *Philosophical Transactions of the Royal Society B*, *373*, 20170121.

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, *143*, 263–292.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Jounral of Artificial Intelligence Research*, *49*, 1–47.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911.

Dove, G. (2015). Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. *Psychonomic Bulletin & Review*, Online First Articles.

Dove, G. (2018). Language as a disruptive technology: Abstract concepts, embodiment and the flexible mind. *Philosophical Transactions of the Royal Society B*, *373*, 20170135.

Gibbs, R. (2006). *Embodiment and cognitive science*. Cambridge University Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

Kiela, D., & Clark, S. (2015). Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2461–2470).

Kiela, D., Hill, F., Korhonen, A., & Clark, S. (2014). Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 835–841).

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14–34.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*, 273–302.

Louwerse, M. M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science*, *10*(3), 573–589.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representation (ICLR)*.

Silberer, C., Ferrari, V., & Lapata, M. (2017). Visually grounded meaning representations. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, *39*(11), 2284–2297.

Takano, K., & Utsumi, A. (2016). Grounded distributional semantics for abstract words. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society (CogSci2016)* (pp. 2171–2176).

Thill, S., Padó, S., & Ziemke, T. (2014). On the importance of a rich embodiment in the grounding of concepts: Perspectives from embodied cognitive science and computational linguistics. *Topics in Cognitive Science*, *6*, 545–558.

Utsumi, A. (2018). A neurobiologically motivated analysis of distributional semantic models. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society (CogSci2018)* (pp. 1147–1152).

Vigliocco, G., Kousta, S.-T., Rosa, P. A. D., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: The role of emotion. *Cerebral Cortex*, *24*, 1767–1777.

Vincent-Lamarre, P., Massé, A. B., Lopes, M., Lord, M., Marcotte, O., & Harnad, S. (2016). The latent structure of dictionaries. *Topics in Cognitive Science*, *8*, 625–659.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464.