OrBEAGLE: Integrating Orthography into a Holographic Model of the Lexicon

George Kachergis, Gregory E. Cox, and Michael N. Jones

Indiana University, Bloomington, IN 47405 U.S.A.
{gkacherg, grcox, jonesmn}@indiana.edu

Abstract. Many measures of human verbal behavior deal primarily with semantics (e.g., associative priming, semantic priming). Other measures are tied more closely to orthography (e.g., lexical decision time, visual word-form priming). Semantics and orthography are thus often studied and modeled separately. However, given that concepts must be built upon a foundation of percepts, it seems desirable that models of the human lexicon should mirror this structure. Using a holographic, distributed representation of visual word-forms in BEAGLE [12], a corpus-trained model of semantics and word order, we show that free association data is better explained with the addition of orthographic information. However, we find that orthography plays a minor role in accounting for cue-target strengths in free association data. Thus, it seems that free association is primarily conceptual, relying more on semantic context and word order than word form information.

Keywords: holographic reduced representation, orthographic priming, semantic priming, word association norms

1 Introduction

Verbal behavior is a hallmark of humankind, and is thus of great interest to cognitive science. Human adults have command of tens of thousands of words and use them effortlessly each day. As such, words are used in studies of many levels of cognition, from perception to memory to semantics, and many effects have been observed. Descriptive factors such as word frequency, length, part-of-speech, and phonological features have been found to be correlated with these effects, but models have rarely attempted to integrate all of these dimensions. Models of word perception tend to focus on the orthographic and phonological features of a word, yet often ignore semantic information such as word co-occurrence. On the other hand, most models of semantics treat words as atomic units with no overt perceptual features. Fortunately, recent research seeks to bridge this divide from both directions.

The SOLAR model [6, 5] uses a spatial coding representation of word forms to account for effects in both masked priming and lexical decision. [3] proposes an alternative orthographic encoding scheme that better captures a set of empirical constraints enumerated by [9]. This encoding scheme uses Holographic Reduced Representations (HRRs) [15], which are neurally-plausible, distributed,

and which result in analytically similar representations for items with similar content or structure. [3] integrates orthographic vectors into BEAGLE, an HRR-based model of lexical semantics and word order [12], and also successfully accounts for lexical decision data. In the present work, after briefly describing HRRs, the Cox et al. orthographic encoding scheme, and BEAGLE, we apply this model to human free-association data [4].

2 Methodology

2.1 Holographic Reduced Representations

HRRs [15] are a form of distributed representation that can hierarchically encode information from diverse sources in a single format. This format is usually a large vector, similar to a layer in an artificial neural network. As in a neural network, it is not the individual vector elements that carry information, but the pattern of values across the vector. The high-dimensional, distributed nature of HRRs thus makes them robust against input noise and memory degradation. Further tightening the connection between HRRs and neural systems, [10] have recently shown that back-propagation neural networks, when trained on location-invariant visual word recognition, produce patterns of stimulus similarity that are equivalent to those derived from HRR representations of the stimuli. HRRs go beyond many simple neural systems, however, in that they can perform variable binding operations (i.e., tracking which items have what properties) and straightforwardly encode hierarchical structure [15].

Although there are other ways of implementing HRRs (including, e.g., binary spatter codes [13]), we focus on the methods introduced by [15] that are based on circular convolution. HRRs begin with a set of "atomic" vectors which are operated upon to produce more structured representations. Each element of these "atoms" is drawn independently from a normal distribution with mean 0 and variance $\frac{1}{n}$, where n is the dimensionality of the vector. There are two operations that enable these atoms to be combined into more structured representations. The first, *superposition* ($+$), is simply vector addition; it takes two HRR vectors and produces a third vector–still an HRR, and with the same dimensionality–that is partially similar to its components (where "similarity" is defined below).

The second operation, *binding* ($\circledast$), takes two HRRs and produces a third HRR that is independent of (not similar to) its components. Binding is implemented as circular convolution, which is both neurally plausible [8] and approximately invertible via correlation ($\#$). If $C = A \circledast B$ is the circular convolution of two vectors, A and B, then each element c_j of C is defined:

$$c_j = \sum_{k=0}^{n-1} a_k b_{j-k \bmod_n}.$$

C can be thought of as a compressed version of the outer product of A and B. Note that the output vector of a circular convolution is the same dimensionality

as each input vector, unlike techniques in other models that produce outputs with greater dimensionality (e.g., [14, 11]). Circular convolution is commutative, associative, and distributes over addition. Implementing circular convolution as defined above is an $O(n^2)$ operation; therefore, in our implementation, we employ the fast Fourier transform, which can be used to approximates circular convolution in $O(n \log n)$ time[1]. In combination, binding and superposition can be used to implement a variety of encoding schemes that simultaneously represent structure at multiple levels. For example, the word *cat* may be represented as the superposition of bound substrings of the word, e.g.: $c+a+t+c \circledast a+a \circledast t$, where each letter is represented by a unique random vector (i.e., they are the "atoms" of the representational scheme). This strategy of chunking long sequences (e.g., letters in words, words in sentences) allows the representation to capture similarity at many resolutions: *cat* will be similar to *catch*, but *catcher* will be more similar to *catch* by virtue of more shared substrings. The similarity between two HRRs is given by their normalized dot product, otherwise known as *cosine similarity*:

$$\mathrm{sim}\,(A, B) = \frac{A \bullet B}{\|A\|\|B\|} = \frac{\sum_{i=0}^{n-1} a_i b_i}{\sqrt{\sum_{i=0}^{n-1} a_i^2}\sqrt{\sum_{i=0}^{n-1} b_i^2}}.$$

This similarity measure is always in the range $[-1, 1]$. The expected cosine similarity of two i.i.d. random vectors (e.g., letters c and a) is 0—that is, they are orthogonal. Bound items (e.g., $c \circledast a$) are independent of (orthogonal to) their contents (c or a), but superposed vectors (e.g., $c + a$) will have positive similarity to each component. Identical vectors have maximal similarity. The similarity of two HRRs relies not just on the contents of the representations (e.g., *cat* and *catch* both have the letters c, a, and t), but also on the structure of the stored associations (e.g., *cat* can be made more similar to *cut* if the association $c \circledast t$ is included in their HRRs). When using HRRs, researchers must be explicit about what structures they are encoding, allowing simpler interpretation and comparison than the learned correlations in standard neural network models.

2.2 A Holographic Encoding for Word-Form

[3] and [9] investigate several methods of encoding word-form structure as a HRR, evaluating them on the basis of empirical studies of word-form similarity. While all but one of these word-form encoding methods were found unable to account for the entirety of the empirical constraints, [3] introduced a word-form encoding that satisfied the desiderata. Our solution, called "terminal-relative" (TR) encoding, is related somewhat to the simplified word recognition model of [1] and to the SERIOL model [16].

Each individual letter is an "atom" represented by a random vector of dimension n with elements drawn independently from a normal distribution $\mathcal{N}\left(0, \frac{1}{\sqrt{n}}\right)$.

[1] This follows from the fact that convolution in the "spatial domain" (i.e., of the raw vectors) is equivalent to elementwise multiplication in the "frequency domain" (the discrete Fourier transforms of each operand).

Thus, the representations for individual letters are orthonormal. To encode a word, e.g., "word", we first superpose vectors for each individual letter and for all contiguous letter bigrams in the word: $word = w+o+r+d+w \circledast o+o \circledast r+r \circledast d$. Here, we wish to form bigrams that are order-specific; to do this, we randomly permute each operand before convolving them according to whether it is on the left or right: $L(w) \circledast R(o)$. To encode larger n-grams, these permutations are applied iteratively: $wor = L(L(w) \circledast R(o)) \circledast R(r)$. Throughout the remainder of this paper, we will use this non-commutative variant of circular convolution (suggested by [15]), although we omit the L and R operators for clarity.

After encoding the individual letters (unigrams) and bigrams, for any n-gram that does not contain one of the terminal letters (either the first or last letter), we encode an additional n-gram that binds the missing terminal letter to that n-gram, including a "space" (just another random vector) to encode a gap in any non-contiguous n-grams. For example,

$$\begin{aligned} word =& w + o + r + d + w \circledast o + o \circledast r + r \circledast d \\ &+ w \circledast o + (w \circledast \_) \circledast r + (w \circledast \_) \circledast d + (w \circledast o) \circledast r + ((w \circledast \_) \circledast r) \circledast d \\ &+ (w \circledast \_) \circledast d + (o \circledast \_) \circledast d + r \circledast d + ((w \circledast o)\_) \circledast d + (o \circledast r) \circledast d \end{aligned}$$

Because this last rule is applied iteratively, the first and last bigrams, and the noncontiguous bigram comprising the terminal letters are added into the representation repeatedly, thus increasing their "strength". Although our method possesses the advantage of being parameter-free, the relative weighting of bigrams in TR encoding is similar to the bigram weighting that arises from neural mechanisms in the SERIOL model of word recognition [16].

The overall effect of this encoding scheme is to capture both local (contiguous bigrams) and global (internal n-grams bound to terminal letters) structure in word-forms. Empirical studies of word recognition (see, for a review, [9, 7]) show that humans are indeed sensitive to structure on both those levels, and that such sensitivity is required to account for human word recognition capabilities. In addition, TR encoding is capable not just of capturing the relative similarity between isolated pairs of words, but scales to account for orthographic similarity effects within the entire lexicon, as evidenced in lexical decision and speeded pronunciation tasks [3]. In general, TR encoding is a good balance between simplicity (it is parameter free), veracity (it accounts for many word recognition effects), and scalability (orthographic similarity effects across the entire lexicon).

2.3 BEAGLE

BEAGLE (Bound Encoding of the AGgregrate Language Environment) is a convolution-based HRR model that learns word order and meaning information from natural language text corpora [12]. For each word, BEAGLE uses an i.i.d. 2048-dimensional *environmental* vector to represent the word's perceptual characteristics, a *context* vector to store word co-occurrence information, and an *order* vector to encode which words appear before and after the given word. As BEAGLE reads each sentence, the environmental vectors of the neighboring

n (window size) words are superimposed on each word's context vector. Words that have similar meanings grow more similar to one another, since their context vectors tend to hold the same set of superimposed vectors. Thus, BEAGLE learns a semantic space; the semantic similarity of any two words can be found by taking the cosine similarity of the two words' context vectors. BEAGLE learns word order by binding n-grams containing a placeholder vector for the current word (denoted Φ) and superimposing the bound n-grams in the current word's order vector. For example, "dog bites" would be encoded in the order vector for "dog" (o_{dog}) as $\Phi \circledast e_{bites}$ where e_{bites} is the environmental vector for "bites". Thus, an approximate environmental vector for the word(s) following "dog" can be obtained by inverting the convolution via the correlation operator (#), $\Phi \# o_{dog} \approx e_{bites}$; we refer to this inversion of the order vector as "probing".

BEAGLE captures many aspects of syntactic and semantic similarity between words. However, this space is constructed on the basis of random environmental vectors which are, on average, orthogonal to one another. We replaced BEAGLE's random environmental vectors with the TR HRR word-form encoding defined above. In this way, we may capture orthographic similarity (e.g., *cat* and *catch*), and perhaps additional semantic relationships (e.g., *catch* and *catcher*). Because OrBEAGLE builds its knowledge of semantics and word order on the basis of a principled orthographic representation, it may better explain a variety of human data, and can be applied to tasks such as fragment completion that no semantic model has previously been suited to model [3]. In the present paper, we examine how well OrBEAGLE accounts for human free-association data, and compare it to BEAGLE.

3 Experiment

[4] collected free association (FA) data by asking some 6,000 participants to write down the first word (*target*) that came to mind after seeing a *cue* word. Given 5,019 words as cues, participants produced 72,176 responses. In various cases, these responses seem to depend on order (e.g., *aluminum-foil*), semantics (e.g., *aluminum-metal*), or orthography (e.g., *abduct-adduct*). Thus, we chose to examine whether OrBEAGLE—which encodes order, semantic, and orthographic information—can account for this FA data.

As a dependent measure, we use the forward strength (FSG) of each cue-target association, defined as the proportion of participants who generated a particular target when given the cue word ($\Pr(\text{target}|\text{cue})$). For example, 17 of 143 participants responded *capability* when given the cue *ability*, so FSG for this pairing is 0.119. We examine how well $logit(\text{FSG})$[2] of each cue-target pairing is predicted by the cosine similarity of these words' representations in BEAGLE and OrBEAGLE.

Using a window size of three and 2048-dimensional vectors, we trained both OrBEAGLE and BEAGLE on the lemmatized TASA corpus (≈680,000 sen-

[2] Because FSG is a probability, the *logit* function is used to transform the domain to all real numbers.

tences). Overall, the cosine similarity of the composite OrBEAGLE representation–including semantic context, order, and orthographic information–of cues and targets are significantly correlated with FSG ($r = .087$, $p < .001$). However, the cue-target similarities computed from BEAGLE's context and order vectors were more strongly correlated with FSG ($r = .199$, $p < .001$). By examining the separate order, context, and environmental (random or orthographic) vectors comprising BEAGLE and OrBEAGLE, we found that context vectors built from BEAGLE's orthogonal environmental vectors were more highly-correlated with FSG than OrBEAGLE's context vectors ($r = 0.154$ vs. $r = 0.035$).

To determine whether OrBEAGLE can explain any unique variance beyond BEAGLE, we used a linear regression to examine the relative contributions of BEAGLE's context and order similarities, and OrBEAGLE's orthographic similarities. Also included in the regression are the cosine similarities of the target's environmental vector with the results from "probing" the order vector of the cue word as described above. Shown in Table 1 (left), BEAGLE's order, context, and probe similarities are significant positive predictors. Introducing OrBEAGLE's orthographic similarities and probe results significantly increased the regression fit (right; $F(1,47858) = 15.687$, $p < .001$). All correlation coefficients are significantly positive, with context the largest, followed by the probe, orthography, and order. A scatterplot of the cue-target similarities and the probe results used in this regression, along with *logit*(FSG), are shown in Figure 1.

Table 1. Regression terms and coefficients (βs) for predicting *logit*(FSG) on the basis of cosine similarities of a cue's and target's holographic vectors, both in ordinary BEAGLE (left), and using OrBEAGLE's orthography and probe results (right).

	BEAGLE			BEAGLE + Ortho.		
Predictor	β	t-value	p-value	β	t-value	p-value
Ortho.	–	–	–	0.249	3.961	<.001***
Context	0.846	28.142	<.001***	0.848	28.188	<.001***
Order	0.207	8.330	<.001***	0.206	8.306	<.001***
Probe	0.598	5.079	<.001***	0.605	5.140	<.001***

4 Discussion

We have described OrBEAGLE, a holographic model incorporating orthographic, semantic, and order information, and demonstrated that it can account for significant variance in human free-association data. However, using independent representations for each word, BEAGLE better accounts for this data, and does so primarily due to its context vectors. In different tasks, perceptual (orthography) and conceptual (context) information likely contribute differentially. Masked priming, a perceptual task, shows large effects of orthography. Lexical decision and word naming are less well-accounted for by orthography, and it

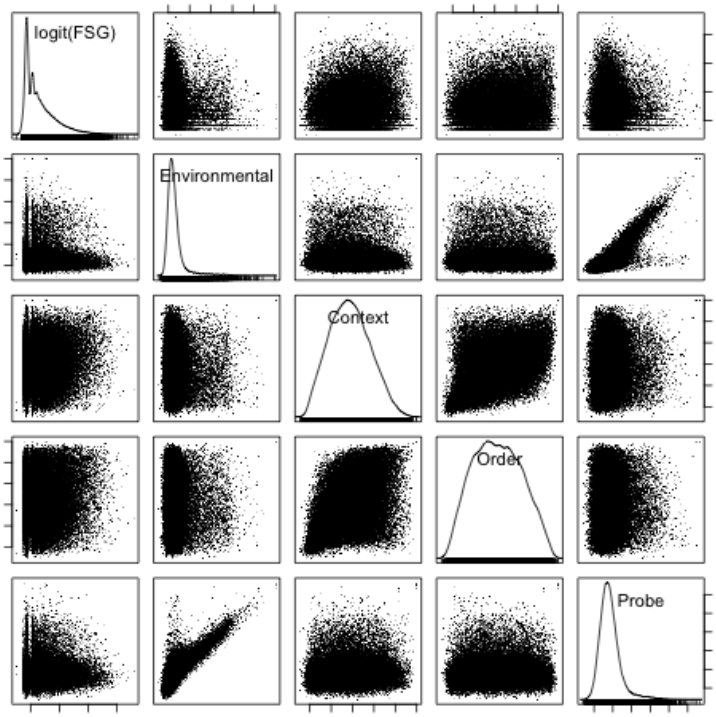

Fig. 1. Cue-target similarities of BEAGLE's context and order vectors and OrBEAGLE's environmental vector and probe results compared to *logit*(FSG).

is not unreasonable to expect that free association would be primarily conceptual. Nonetheless, we also demonstrated that a significant portion of additional variance in FA data can be accounted for by adding OrBEAGLE's orthographic similarities and orthographic probe results to BEAGLE's context and order similarities.

Indeed, the partial independence of orthographic and semantic/syntactic properties of words underlies many theories of verbal processing, including the Dual Route Cascaded model (DRC, [2]). In DRC, orthographic similarity plays a role in *which* semantic representations are activated, but the two types of information are not embodied in a single representation; DRC's semantic representations are akin to the random environmental vectors in the original BEAGLE model, while its orthographic representations are akin to the environmental vectors in OrBEAGLE. While DRC has been implemented as a neural network model, the results in this paper suggest that a holographic approach—with reduced training time and the ability to store a larger lexicon—can capture many of the same theoretical ideas.

OrBEAGLE is a distributed representation system that uses neurally-plausible mechanisms and an empirically-viable encoding scheme for visual word-forms to

capture a wide variety of human data, including latencies in word naming and lexical decision, word fragment completion [3], and now, free-association data. OrBEAGLE can be applied to a variety of other experimental paradigms as it stands, and can indicate the relative contributions of word order, context, and orthography. We have shown that it is both possible and useful to begin unifying models that previously operated at different levels and on different tasks.

References

1. Clark, J.J., O'Regan, J.K.: Word ambiguity and the optimal viewing position in reading. Vision Res. 39, 843–857 (1998)
2. Coltheart, M., Rastle, K., Perry, C., Langdon, R., Ziegler, J.: The DRC model: A model of visual word recognition and reading aloud. Psychol. Rev. 108, 204–258 (2001)
3. Cox, G.E., Kachergis, G., Recchia, G., Jones, M.N.: Towards a scalable holographic word-form representation. Behav. Res. Methods (in press)
4. D. L. Nelson, C. L. McEvoy, T.A.S.: The university of south florida word association, rhyme, and word fragment norms (1998), http://www.usf.edu/FreeAssociation/
5. Davis, C.J.: The spatial coding model of visual word identification. Psychol. Rev. 117(3), 713–758 (2010)
6. Davis, C.J.: The self-organising lexical acquisition and recognition (SOLAR) model of visual word recognition. Ph.D. thesis, University of New South Wales, Sydney, Australia (1999)
7. Davis, C.J., Bowers, J.S.: Contrasting five different theories of letter position coding: Evidence from orthographic similarity effects. J. Exp. Psychol. Human 32(3), 535–557 (2006)
8. Eliasmith, C.: Learning context sensitive logical inference in a neurobiological simulation. In: Levy, S., Gayler, R. (eds.) Compositional connectionism in cognitive science. AAAI Press, Menlo Park, CA (2004)
9. Hannagan, T., Dupoux, E., Christophe, A.: Holographic string encoding. Cognitive Sci. 35(1), 79–118 (2011)
10. Hannagan, T., Dandurand, F., Grainger, J.: Broken symmetries in a location invariant word recognition network. Neural Comput. (in press)
11. Humphreys, M.S., Bain, J.D., Pike, R.: Different ways to cue a coherent memory system: A Theory for episodic, semantic, and procedural tasks. Psychol. Rev. 96(2), 208–233 (1989)
12. Jones, M.N., Mewhort, D.J.K.: Representing word meaning and order information in a composite holographic lexicon. Psychol. Rev. 114(1), 1–37 (2007)
13. Kanerva, P.: The spatter code for encoding concepts at many levels. In: P. Int. Conf. Artif. Neural Networ., vol. 1, pp. 226–229. Springer-Verlag, London (1994)
14. Murdock, B.B.: A theory for the storage and retrieval of item and associative information. Psychol. Rev. 89(3), 609–626 (1982)
15. Plate, T.A.: Holographic Reduced Representations. CSLI Publications, Stanford, CA (2003)
16. Whitney, C.: How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. Psychon. B. Rev. 8(2), 221–243 (2001)