LANGUAGE, COGNITION AND NEUROSCIENCE, 2018
https://doi.org/10.1080/23273798.2018.1431679

REGULAR ARTICLE

When does abstraction occur in semantic memory: insights from distributional models

Michael N. Jones

Michael Jones Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA

ABSTRACT

Abstraction is a core principle of Distributional Semantic Models (DSMs) that learn semantic representations for words by applying dimensional reduction to statistical redundancies in language. Although the posited learning mechanisms vary widely, virtually all DSMs are prototype models in that they create a single abstract representation of a word's meaning. This stands in stark contrast to accounts of categorisation that have very much converged on the superiority of exemplar models. However, there is a small but growing group of accounts in psychology, linguistics, and information retrieval that are exemplar-based semantic models. These models borrow many of the ideas that have led to the prominence of exemplar models in fields such as categorisation. Exemplar-based DSMs posit only an episodic store, not a semantic one. Rather than applying abstraction mechanisms at learning, these DSMs posit that semantic abstraction is an emergent artifact of retrieval from episodic memory.

ARTICLE HISTORY
Received 2 December 2016
Accepted 17 January 2018

KEYWORDS
Semantic memory; latent semantic analysis; categorisation; abstraction

Abstraction is an essential mechanism to learn and represent meaning in memory. Theoretical notions of abstraction vary across research domains, but tend to emphasise aggregation across exemplars to a central "average" representation (Reed, 1972), transforming sensorimotor input to a deeper knowledge representation (Barsalou, 1999; Damasio, 1989), or reducing idiosyncratic dimensions to focus on those attributes most common to members of a category (Rosch & Mervis, 1975). In modern computational models of semantic memory, notions of abstraction are formally specified and applied to real-world linguistic data to evaluate the structure of semantic memory that the mechanisms would produce.

Modern distributional semantic models (DSMs; e.g. Landauer & Dumais, 1997) have become immensely popular in the cognitive literature due to their success at fitting human experimental data, their utility in real-world applications, and their insights as models of cognition. In general, DSMs learn distributed representations for word meanings from statistical redundancies across linguistic experience. Because they are often applied to text corpora as learning data, DSMs are also referred to as "corpus-based" models, although, in principle, their learning mechanisms can be applied to covariational structure in any dataset (e.g. perception, speech, etc.).

Despite the wide range of DSMs in the literature, they virtually all share the characteristic that they are prototype models: They attempt to collapse the entire set of a word's linguistic exemplars into a single economical representation of word meaning. However, this practice is in contrast to the literature on categorisation that has largely disposed of prototype representations in favour of exemplar-based models. In this paper, I highlight the contradiction between literatures, and attempt to build a case for exemplar-based models of distributional semantics.

Abstraction is a core mechanistic principle of DSMs. Most DSMs apply some form of dimensional reduction to words' experienced linguistic contexts, essentially abstracting over the dimensions that are idiosyncratic to each context, and converging on the stable higher-order dimensions that optimally explain the covariational pattern of words across contexts. Aggregation is also a core principle of virtually all DSMs – multiple linguistic contexts are averaged across, either explicitly or implicitly, resulting in a single central representation for the word that is stored. A word's vector pattern across these reduced dimensions is thought to represent its generic meaning. Hence, each DSM formally specifies an abstraction mechanism by which episodic memory is transformed into semantic memory; in this sense, DSMs embody the idea of abstraction and allow us to quantitatively evaluate various process explanations of aggregation and dimensional reduction.

CONTACT Michael N. Jones ✉ jonesmn@indiana.edu

In Press: *Language, Cognition, and Neuroscience*, Special Issue on "Concepts: What, When, and Where".

The particular mechanisms posited for abstraction in DSMs differ in several theoretically important ways, and include reinforcement learning, probabilistic inference, latent induction, and Hebbian learning. Enumerating the differences between the mechanisms used by each model is beyond the scope of this article (see Jones, Willits, & Dennis, 2015 for a review); but all DSMs essentially specify an abstraction mechanism to formalise the classic notion in linguistics that "you shall know a word by the company it keeps" (Firth, 1957). The theoretical points that follow apply broadly to all DSMs that posit abstraction *at learning*, regardless of specific learning mechanism. As two examples of DSMs with very different architectures and learning mechanisms,[1] briefly consider classic Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) and the newest DSM – Google's word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013).

LSA begins with a word-by-document frequency matrix of a text corpus. This initial "episodic" matrix represents first-order relationships: words are similar if they have frequently co-occurred in contexts. LSA then applies singular value decomposition to this episodic matrix (cf. factor analysis) retaining only the 300 or so dimensions that account for the largest amount of variance in the original matrix. Singular value decomposition serves as an abstraction mechanism, reducing the dimensionality and emphasising second-order relationships that were not obvious in the episodic matrix. In the reduced space, words will be similar if they occur in similar contexts, even if they never directly co-occur (e.g. category exemplars and synonyms). But much of the information idiosyncratic to specific contexts that would be required to reconstruct the full original episodic matrix has now been lost. Hence, LSA achieves an abstracted semantic representation by applying truncated SVD to the history of episodes.

Mikolov et al.'s (2013) word2vec achieves a similar outcome, albeit in a rather different way. Word2vec is a "neural embedding" model that has been extremely successful in computational linguistics. To a cognitive scientist, the model is essentially a feedforward connectionist network (cf. Rumelhart networks explored in Rogers & McClelland, 2004) with some optimisation tricks that allow it to be scaled up to large amounts of text data. Word2vec has localist input and output layers, each with one node for each word in the corpus. The input and output layers are fully connected via a hidden layer of ~300 nodes, which allows the model to learn nonlinear patterns in the text corpus. When a word is experienced, the other words that it occurs with serve as its context. With the node for the context words activated, activation feeds forward to the output layer with the desired output being the activation of the correct target word, with other words being inhibited. The error signal (difference between true and observed output pattern) is then backpropagated through the network to increase the likelihood that the correct word will be activated at the output layer given the input words in the future. Hence, the context is used to predict the word.[2] After training on a large text corpus, a word's pattern across the hidden layer begins to show higher-order relationships that go beyond the first order relationships it was being trained to predict. Very much like LSA's reduced representation, the reduced representation across word2vec's hidden layer has now learned similarity between words that are predicted by similar contexts. While LSA used SVD for data reduction, word2vec used backpropagation; but both models essentially abstract semantics from episodes.

These similarities can be seen across all of the DSMs – all achieve the desired outcome of a reduced abstraction of word meaning from episodic co-occurrences. The jury is still out on which (if any) mechanism is the most plausible model of how humans construct semantic representations. But one property that is clear to all of these "abstraction at learning" DSMs is that they may be classified as *prototype* models. The models attempt to create a single abstracted representation of meaning for each word, and this single semantic representation is what is stored and used in downstream fitting of psycholinguistic data.

There are many similarities in the literatures on semantic memory and categorisation, enough that it is likely that the cognitive mechanisms that subserve semantic learning and category learning may be heavily related to each other. But one key contradiction stands out: While the literature on categorisation has very much converged on the superiority of exemplar-based models, DSM models are all essentially prototype models.

Lessons from categorization models

Categorisation and semantic abstraction have many similarities, and it is commonly believed that the process of categorisation may be used to produce semantic structure (see Rogers & McClelland, 2011 for a review). The categorisation literature has been dominated for many years by a debate between prototype and exemplar-based theories. Prototype theories are based largely on principles of *cognitive economy* championed by Rosch and Mervis (1975). Prototype theories (e.g. Reed, 1972) posit that as category exemplars are experienced, humans gradually abstract generalities across them and construct a single prototypical representation of the category that is the central tendency of its exemplars; categorisation of a new exemplar depends on its similarity to category prototypes. In contrast, exemplar theories (e.g. Medin & Schaffer, 1978; Nosofsky, 1988) posit that humans store every experienced exemplar in memory, and categorisation of a new exemplar depends on its weighted similarity to all stored exemplars.

Perhaps more than any other sub-field of cognition, the categorisation literature has very much converged on the conclusion that exemplars have beaten out prototypes as models of human categorisation (but see Murphy, 2016, for a careful discussion of the limitations of both). In addition to exemplar models providing a better quantitative account of human categorisation data, there are many theoretically differentiating effects that are easily explainable by exemplar models but that are simply impossible under prototype accounts. For example, category structures with nonlinearly separable structure (e.g. XOR) are easily learned by humans, but impossible to account for by single prototype models (Ashby & Maddox, 1993; Nosofsky, 1988). Even when using linear category structures that should be conducive to prototype models, exemplar models still give a superior quantitative fit to human data (Stanton, Nosofsky, & Zaki, 2002). Hence, it is certainly odd that the field of distributional semantics is dominated by prototype models, while the field of categorisation has largely dismissed them in favour of exemplar accounts.

In the typical categorisation experiment, subjects are presented with stimulus patterns – exemplars – accompanied by a category label. At test, the experimenter can present old or new exemplars, and the subject responds with the most appropriate category label for each stimulus. We can think of distributional learning of semantics in an analogous way: The context is the exemplar pattern, and the word is the label of the category to which this particular exemplar belongs.

In word2vec, for example, the other words that occur with a target word are used as the context, or exemplar pattern, and the correct label is the target word. So in the sentence "I am drinking a glass of milk," *drinking* + *glass* are used as the context to predict *milk*. The exemplar pattern for *milk* in this context is a localist vector with *drinking* and *glass* set to one and all other words set to zero. Across many language exemplars that are all of the category *milk*, word2vec homes in on a pattern of activation across its hidden layer that optimally predicts *milk* as the label given any language exemplar context that contains *milk*. In addition, the hidden layer pattern for *milk* will be very similar to other words that are predicted by similar contexts such as *juice* and *wine*. So in all DSMs, an exemplar can be thought of as the context pattern of other words that a target word (the category label) occurs with. This reframing of semantic learning is very similar to current state-of-the-art exemplar-based models of categorisation in which " … a stimulus is stored in memory as a complete exemplar that includes the full combination of other features. Thus the 'context' for a feature is the other features with which it co-occurs." (Kruschke, 2008, p. 273). However, DSMs aggregate over the multiple exemplars to create an economical prototype.

Hence, most DSMs collapse all instances of a word's context into a single representation, or point in high-dimensional space, very much consistent with representational economy (Rosch, 1973). This process produces huge issues in semantic representation that are known to the field – for example, a homograph like *bank* has both senses of its meaning collapsed into a single representation, despite the fact that they are very different context patterns. As a result, the representation becomes a weighted average (biased to the more frequent sense) of the multiple senses of *bank*. A homograph like *bank*, with multiple unrelated senses, has a similar characteristic structure to experimental stimuli with XOR structure. But the prototype collapsing is a problem for all words with graded amounts of polysemy that would be captured by an exemplar-based model but are abstracted over by a prototype-based DSM. Multiple distinct statistical structures that map onto the same label are collapsed in most DSMs, leading to a range of both theoretical and practical issues for the models. But far from rare, multiple senses and contextual modulation patterns are really the norm in linguistic information (Jones, Dye, & Johns, 2016; Kintsch, 2001).

Lessons from multiple-trace models

Posner and Keele's (1968) schema abstraction experiment is a classic in semantic memory research, and was a key laboratory phenomenon that lead Tulving (1972) to divide declarative memory into separate semantic and episodic stores in his modular taxonomy.[3] In their task, Posner and Keele presented subjects with random dot patterns as exemplars of multiple categories. Unbeknownst to subjects, the exemplars they experienced were created from parent prototype patterns for each category. A category exemplar was a random perturbation (low or high distortion) of the prototype pattern, but prototypes were never shown to subjects during learning. There are many interesting effects from the schema abstraction task, but a key finding is that while subjects at test are better at classifying exemplars they were trained on, they were better at the prototype than new exemplars. In addition, with a delay between training and test, performance on the prototype (which was never experienced) is *better* than performance on the old exemplars that subjects were actually trained on. Furthermore, exemplars and prototypes follow different trajectories of decay as a function of retention time.

This pattern of results suggests that subjects are storing experienced exemplars in episodic memory at the same time as they are creating an abstracted prototype for the category. The differential decay patterns also suggest that these information sources are stored by distinct memory systems, and that semantic memory is more resilient to decay than is episodic memory. The pattern would seem to argue in favour of DSMs that use abstraction at encoding to create a prototypical pattern, and episodic memory is then explained by a distinct model.

However, Hintzman (1984; 1986) provided a classic demonstration using his MINERVA 2 memory model that questioned whether the effects seen in Posner and Keele's (1968) schema abstraction task suggest the existence of a prototype in memory at all. Briefly, MINERVA 2 is a instance-based memory model: it stores a pattern for each exemplar in episodic memory, but has no semantic memory. Multiple presentations of an exemplar simply lay down multiple memory traces. The model explains a range of episodic memory effects such as recognition, judgments of frequency, etc. But it can also perform the classification task used in the schema abstraction experiments. When presented with a probe pattern (an old or new exemplar) MINERVA 2 simultaneously computes the probe's similarity to all stored exemplar traces in memory, and the retrieved category label for the probe is weighted by the scaled similarity of the probe to all exemplars (cf. Nosofsky's, 1986 exemplar-based model of categorisation). The retrieved pattern is referred to as an "echo" from memory, and is based loosely on the principle of harmonic resonance. Although it has no semantic memory per se, MINERVA 2 reproduces the key phenomena in schema abstraction that had previously been seen as evidence for dual episodic and semantic stores. The model performs better on old exemplars at immediate test (but better on the prototype than new exemplars), and performance on the prototype is better than the training exemplars after forgetting. Superior performance on the prototype is due simply to the fact that it is the central tendency of the exemplar patterns; hence the prototype's pattern is distributed across the exemplars. The performance trajectories of exemplars and the prototype as a function of delay have distinct slopes.

Hintzman's (1984; 1986) demonstration is well covered in most contemporary memory textbooks – it is an elegant existence proof that phenomena used to argue for the existence of semantic memory may actually be due to the process of retrieval from episodic memory. In the interest of parsimony, there may be no need to posit an additional semantic store when a model that has only an episodic store can produce all the phenomena that a model with two distinct stores could. This claim bears considerable similarity to other instance-based models of memory and exemplar-based models of categorisation. So why mention historical cases like MINERVA 2 and schema abstraction here? Because one of the first successful exemplar-based DSMs in cognitive science extends MINERVA 2's architecture exactly to a text corpus, and makes the same theoretical claims.

Exemplar-based semantic models

While it is true that most DSMs are prototype models, there is a small family of exemplar-based semantic models that diverge from the usual quest for cognitive economy. Exemplar-based semantic models are also referred to as "retrieval-based" models in the cognitive literature or simply as "memory models" in computational linguistics. Rather than positing abstraction as a dimensional reduction mechanism at learning, they store all of a word's episodic contexts, and abstraction is a consequence of retrieval from episodic memory. Hence, there is *no semantic memory* per se in these models, only episodic memory. In exemplar-based models, phenomena that have typically been attributed to semantic memory are an emergent artifact of retrieval from episodic memory. The locus of semantics is not at encoding, but at retrieval. These models have grown from exemplar-based models in categorisation, and instance-based models in memory. Intuitively, many people believe the idea that we store everything we ever experience rather than creating and storing an economical abstraction is far-fetched. But given the success of exemplar-based semantic models at accounting for an impressive array of semantic behaviours without any semantic memory, and the current resurgence of usage-based theories in linguistics (Goldberg, 2006; Johns & Jones, 2015; Tomasello, 2003), exemplar-based semantic models deserve a closer look.

Kwantes (2005) extended Hintzman's (1986) MINERVA 2 to explain semantic phenomena with words by training it on a text corpus. In his Constructed Semantics Model (CSM),[4] each word's representation in memory is a binary vector that reflects whether it occurred in a document or not – its episodic history. Note that memory in CSM is the same word-by-document matrix that LSA and other DSMs learn from. But where LSA applies abstraction to this episodic matrix and stores a higher-order representation, CSM stores the episodic matrix itself. When a word is presented to CSM, its episodic vector is used as a probe as in MINERVA 2. Each word in memory is activated relative to its contextual overlap with the probe word, and the echo pattern is then the similarity-weighted sum of all traces in memory, exactly as in Hintzman (1986). Words that have similar contextual histories to the probe word will contribute more of their pattern to the echo than will words with rather independent histories, and the echo for the word is then an ad hoc, and probe specific, prototype created by the this process of retrieval from episodic memory. To compute the semantic similarity between two words, one simply computes the cosine between their two echo patterns.

It is fairly obvious how CSM can determine similarity for words that frequently co-occur with each other (their echo cosines would be a noisy amplification of the terms' likelihood of co-occurrence relative to chance). But higher-order semantic similarities also emerge from this process of retrieval, even between two words that have zero contextual overlap. For example, two synonyms would not activate each other at all because they have never co-occurred in the text corpus, but they would activate many of the same other words due to their similar contextual usage; as a result, their retrieved echo patterns are extremely similar. Models such as LSA and word2vec accomplish this second-order statistical inference while learning a corpus, whereas CSM does it while retrieving information from episodic memory.

Hence, semantic abstraction in CSM is a parallel to schema abstraction in MINERVA 2: the prototype is an emergent property of retrieval from episodic memory. As Kwantes (2005) puts it, CSM " … takes what it knows about a word's contexts and uses retrieval to estimate what other context might also contain the word" (p. 706). The model bears obvious similarity in outcome to prototype-based DSMs, but it differs considerably in the psychological mechanism that it attributes abstraction to; in CSM, it is the well-established process of retrieval that uncovers deeper semantic structure.

As with Hintzman's (1986) demonstration, CSM is a more parsimonious model of semantics – it does not require two separate stores or processes to explain semantic and episodic memory, and serves as an existence proof that semantic phenomena may be explained by a model that only posits an episodic memory store. In addition, the success of the model is reinforced by converging evidence supporting exemplar-based models in the fields of categorisation and recognition. Furthermore, there are real benefits to CSM that allow it to handle phenomena not possible by abstractionist DSMs. For example, it can handle polysemous words because the multiple senses of the words are still represented and are dissociable with nonlinear activation of exemplars (cf. Nosofsky, 1986). Memory traces whose context fits one or the other sense of a word can be differentially activated in CSM. Abstractionist DSMs, on the other hand, collapse multiple senses of a word to a single point in high-dimensional space, losing the distinction in favour of an averaged representation.[5]

Kwantes (2005) work suggests that the same basic memory system could underlie both episodic and semantic knowledge, and his work has given rise to a handful of other models that have explored semantic abstraction as a memory retrieval operation rather than a learning mechanism. For example, Dennis (2005; see also Thiessen, 2017) presented a memory-based model of verbal processing, including semantics and syntactic information as retrieval from long-term memory and constraint satisfaction in working memory. The model mechanisms are based on a Bayesian interpretation of string edit theory from linguistics. Dennis' model posits that processing a word or sentence is at its core a memory-retrieval process.

Johns and Jones (2014, 2015; see also Thiessen & Pavlik, 2013) extended this previous work into an exemplar-based model, based on a hybrid of Hintzman's MINERVA 2 (1986) and Jones and Mewhort's (2007) BEAGLE architectures, that encodes sentences from a natural language text corpus into individual memory exemplars. The retrieval mechanism is used to generate expectancies about the future structure of sentences, much in the same way as Kwantes (2005) constructs a word's meaning as a prediction of the future contexts in which it might occur. Johns and Jones found that such an exemplar-based model successfully accounted for a wide range of sentence processing tasks that had commonly been seen as evidence for rule-based abstraction of linguistic constraints. Johns, Jamieson, Crump, Jones, and Mewhort (2016) extended this model to demonstrate that rule-based grammatical behaviour is a natural emergent property of retrieval from a model that stores exemplars of linguistic experience. Hence, both semantics and syntax may very well be constructed properties of retrieval from episodic memory rather than abstracted structures or rules, per se.[6]

Exemplar-based models in natural language processing

It is tempting to think of exemplar-based models as a psychology centric theory with little, if any, practical significance. After all, why would a computing scientist want to store all data instances? Data compression and abstraction are core goals to information retrieval applications. However, exemplar models are now seeing considerable use in natural language processing (NLP) as well, for the very same reasons that they are preferred in categorisation: the affordance of nonlinear activation of memory exemplars given a probe.

A classic example in NLP was presented by Daelemans, Van Den Bosch, and Zavrel (1999), showing that abstractionist models lose exceptions to common patterns in a variety of language processing tasks. Prototype models offer the best single representation, but the distribution of meanings and usage rules is heavily skewed in natural languages – prototype models discard the tail (cf. Johns & Jones, 2010 in lexical semantics). Daelmans et al. found that retaining exceptional training instances in memory was actually beneficial for generalisation accuracy across a wide range of common NLP tagging tasks, and they argue that the field needs to take exemplar-based memory models much more seriously.

More recently, exemplar-based models have seen a resurgence in NLP, offering better accuracy on applied problems that the field had been deadlock on with abstractionist models. For example, Erk and Padó (2010) used an exemplar-based memory model in which separate exemplars were encoded for words and sentences (cf. Dennis, 2005; Johns & Jones, 2015). They found superior performance on a practical paraphrase task using this architecture due to the nonlinear activation of related exemplars – this behaviour allowed the model to "ignore" the exemplars that were other senses of a target word, which would have been a collapsed noise source in an abstractionist model. In fact, their exemplar model outperformed all then state-of-the-art paraphrasing models, and has considerable similarity to exemplar-based memory models in cognitive science (e.g. Thiessen & Pavlik, 2013).

However, an issue with the application of exemplar-based models to applied NLP tasks will always be processing time. Exemplar-based models are fast to train, but require substantial and even computationally impractical memory resources, and are slow to retrieve the correct answer. In contrast, prototype models embody data compression, putting all the time into training the single best representation, but then the search time for a similar instance in memory is much more efficient. In applied problems, such as information retrieval, access time is everything. However, there have been many successful hybrid models emerging in NLP that balance accuracy with generalisation and speed. Multiple prototype models (e.g. Reisinger & Mooney, 2010) have become popular to represent the distinct senses of a word without needing to store all exemplars, and are quite similar to multiple prototype theories of categorisation (Minda & Smith, 2001). Similarly, there has been considerable success in NLP with models that represent words as regions, rather than points, in distributional space (Erk, 2009; Vilnis & McCallum, 2015). These models preserve nonlinear activation of exemplars, but while embedding them in a more reasonable search space with attractor basins. The practice has a similar outcome to setting a threshold on the similarity function in exemplar-based psychological models to reduce the activation of irrelevant items (which is precisely what Kwantes, 2005, model does). This also suggests considerable potential for the application of hybrid rule-and-exception models from human category learning (e.g. Nosofsky, Palmeri, & McKinley, 1994).

Also of interest in practical NLP applications is the recent rise of so-called memory networks (Weston, Chopra, & Bordes, 2015) that have proven very successful at open question answering with complex real-world text materials. Memory networks use a long-term exemplar memory network as a dynamic knowledge base, and have produced state-of-the-art results with difficult tasks such as question answering, summarisation, and text-based inference (Bordes, Usunier, Chopra, & Weston, 2015).

Discussion

Meaning is a fundamental human attribute that permeates all cognition, from low-level perceptual processing to high-level problem solving, and everything in between. Semantic abstraction is what makes us a powerful species – *informavores*. The idea that humans construct and store abstracted semantic representations for concepts is almost sacred in cognitive science. But it is also at odds with conclusions from other areas of cognition, such as categorisation and recognition, which presumably tap aspects of the same cognitive mechanism as semantic learning. And exemplar-based DSMs suggest that, like Hintzman's (1986) demonstration, we might be able to explain all the same semantic phenomena without a semantic memory. According to exemplar-based DSMs, semantic memory is a process, not a structure.

It is tempting to see exemplar-based DSMs as "cheating:" If the model simply stores all data, then it can compute an accurate semantic representation whenever one is needed. But the theoretical claim is profound – it is a frightening proposal that we may not actually have semantic memory. Your interpretation of the words you are reading right now may be constructed on the fly as an artifact of retrieving the visual patterns from episodic memory. Our phenomenology of meaning may be continuously constructed as the interaction between stimuli, episodic memory, and the memory retrieval mechanism that mediates them (Kintsch & Mangalath, 2011). But exemplar-based DSMs should also put us at ease – they provide converging evidence that performance across multiple cognitive domains (e.g. categorisation, recognition, semantics) may be explained by the same unified cognitive principle. Exploring exemplar-based DSMs also has practical considerations for education, where exemplar-based models of categorisation have been successfully applied (Norman, Young, & Brooks, 2007; Nosofsky, 2017). And since humans are both the producers and consumers of linguistic information in all practical NLP tasks, it is also reassuring that the recent findings in NLP may suggest that the best models to serve humans in these tasks bear considerable similarity to the models we believe human cognition has evolved to use.

How might exemplar-based semantic models be implemented in neural hardware? A common criticism against exemplar theory is that it is stranded at the computational level of Marr's hierarchy, but the transition to implementation is untenable. The claim that we simply store everything that we experience seems unintuitive and goes against the core principle of cognitive economy. However, Hintzman (1990) has shown how an exemplar-based memory model such as MINERVA 2 can be easily implemented within a neural network framework. In addition, there is a small body of work attempting to understand and formalise biologically plausible exemplar theories of recognition and categorisation, which have typically pointed to a role for the hippocampus and surrounding medial temporal lobe structures (e.g. Pickering, 1997). Futhermore, Becker's (2005) models cleanly demonstrate that hippocampal coding would give rise to distinct memory representations for highly similar items.

More recent work in categorisation is now focusing on the basal ganglia and striatum as giving rise to the operations needed for exemplar models (see Ashby & Rosedahl, 2017, for a review). Ashby and Rosedahl recently introduced a neural implementation of exemplar theory, in which a key role for the formation of category exemplars is assigned to synaptic plasticity at cortical-striatal synapses. Rather than storing strict exemplars, per se, their model adds nodes and manipulates connectivity between striatal and sensory neurons, achieving the same effect as classic exemplar models. Ashby and Rosedahl show that their neural implementation of exemplar theory is mathematically equivalent to classic exemplar theories such as the General Context Model (Nosofsky, 1986), and makes identical predictions. The work of Ashby & Rosedahl establishes an important equivalence between classic exemplar based models and neural exemplar theories. Not only do exemplar theories provide superior quantitative fits, but increasing biological plausibility also extends predictions to findings from cognitive neuroscience (e.g. Ashby & Valentin, 2016; Hélie, Paul, & Ashby, 2012; Valentin, Maddox, & Ashby, 2014).

The predominance of prototype-based DSMs in the literature may be partially due to Chomskian presumptions in linguistics that the job of the cognitive mechanism is to abstract the rules of a grammar from instances. This abstractionist presumption may have implicitly guided architectural decisions in early DSMs. In addition, the notion of cognitive economy (Rosch & Mervis, 1975) was a guiding principle to models of semantic abstraction. However, both of these theoretical presumptions are currently being revisited given the strength of usage-based theories in linguistics (Tomasello, 2003). But a more likely reason for the preference of prototype-over exemplar-based DSMs in practice is that exemplar models are much more computationally expensive than prototype models. The front-end data compression core to prototype DSMs means that they require far less memory to store, and are far more efficient to use, than exemplar-models.

Development of DSMs in general has benefited from cross-disciplinary interactions with applied fields, such as information retrieval. Put simply, these models are both theoretically informative to cognitive science, and useful for practical NLP tasks. However, the utility of the model should not constrain its theoretical informativeness. Prototype DSMs are needed because they provide an efficient and economical estimate of a word's aggregate meaning. Exemplar-based DSMs contain more information, but the retrieval problem becomes intractable and untenable for practical tasks. Nobody wants to type a search query into Google and have it determine what you mean by activating and weighting exemplars in real time; the time-intensive computation should have been completed and stored long before you type in the query words.

But the constraint of utility in NLP may have had the unwanted effect of guiding our theoretical models of the mind away from exemplar models. There are many differences between the brain and computational databases in how they represent and retrieve information. The search and abstraction processes used in cognition need not be identical to those best for database search. Models of cognition have long assumed that memory exemplars can be activated in parallel, although the code we use to implement this in a model will usually use a loop routine. This is a distinct difference between the two disciplines: Looping through all exemplars is not an efficient method of, for example, word similarity matching, but it may well be the correct model of how humans do it. The constraints of our current computational hardware should not be used as reasons to discard otherwise superior fitting models of human cognition, such as exemplar models.

Notes

1. LSA and word2vec are formally equivalent: Levy and Goldberg (2014) demonstrated analytically how the SGNS architecture of word2vec is implicitly factorizing a word-by-context matrix whose cell values are shifted PMI values.
2. The model's direction can also be inverted, using the word to predict the context (SGNS) rather than using the context to predict the word (CBOW).
3. The bulk of the evidence used by Tulving to argue for distinct semantic and episodic memory systems was from neuropsychological patients.
4. The model is simply referred to as the "semantics model" in Kwantes' (2005) original paper, but "Constructed Semantics Model" has become it's popular name among semantic modelers because semantic representations are constructed on the fly from episodic memory in the model.
5. An exception here is the topic model, which uses conditional probabilities, so it is not subject to metric restrictions of spatial models (e.g., Griffiths, Steyvers, & Tenenbaum, 2007).
6. And essentially the same architecture has been used by Goldinger (1998) to explain "abstract" qualities of spoken word representation from episodic memory retrieval.

Acknowledgements

This work was supported by NSF BCS-1056744 and IES R305A150546. I would like to thank Randy Jamieson, Melody Dye, and Brendan Johns for helpful input and discussions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by National Science Foundation [grant number BCS-1056744] and Institute of Education Sciences [grant number R305A150546].

References

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400.

Ashby, F. G., & Rosedahl, L. (2017). A neural implementation of exemplar theory. *Psychological Review*, *124*(4), 472–482.

Ashby, F. G, & Valentin, V. V. (2016). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science, second edition* (p. in press). New York: Elsevier.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and Brain Sciences*, *22*(4), 637–660.

Becker, S. (2005). A computational principle for hippocampal learning and neurogenesis. *Hippocampus*, *15*(6), 722–738.

Bordes, A., Usunier, N., Chopra, S., & Weston, J. (2015). *Large-scale simple question answering with memory networks* (arXiv preprint arXiv:1506.02075).

Daelemans, W., Van Den Bosch, A., & Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. *Machine Learning*, *34*(1–3), 11–41.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*(1), 25–62.

Dennis, S. (2005). A memory-based theory of verbal cognition. *Cognitive Science*, *29*, 145–193.

Erk, K. (2009). *Representing words as regions in vector space*. Proceedings of the thirteenth conference on computational natural language learning (pp. 57–65). Association for Computational Linguistics.

Erk, K., & Padó, S. (2010). *Exemplar-based models for word meaning in context*. Proceedings of the acl 2010 conference short papers (pp. 92–97). Association for Computational Linguistics.

Firth, J. R. (1957). *A synopsis of linguistic theory* (pp. 1930–1955). Oxford: Blackwell.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. New York: Oxford University Press.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211.

Hélie, S., Paul, E. J., & Ashby, F. G. (2012). A neurocomputational account of cognitive deficits in Parkinson's disease. *Neuropsychologia*, *50*(9), 2290–2302.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Hintzman, D. L. (1990). Human learning and memory: Connections and dissociations. *Annual Review of Psychology*, *41*(1), 109–139.

Johns, B. T., Jamieson, R. K., Crump, M. J. C., Jones, M. N., & Mewhort, D. J. K. (2016). *The combinatorial power of experience*. Proceedings of the 37th meeting of the cognitive science society.

Johns, B. T., & Jones, M. N. (2010). Evaluating the random representation assumption of lexical semantics in cognitive models. *Psychonomic Bulletin and Review*, *17*, 662–672.

Johns, B. T., & Jones, M. N. (2014). *Generating structure from experience: The role of memory in language*. Proceedings of the 35th annual conference of the cognitive science society. Austin, TX: Cognitive Science Society.

Johns, B. T., & Jones, M. N. (2015). Generating structure from experience: A retrieval-based model of language processing. *Canadian Journal of Experimental Psychology*, *69*, 233–251.

Jones, M. N., Dye, M., & Johns, B. T. (2016). Context as an organizing principle of the lexicon. In B. Ross (Ed.), *The psychology of learning and motivationion* (Vol. 67, pp. 239–283). https://doi.org/10.1016/bs.plm.2017.03.008

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37.

Jones, M. N., Willits, J. A., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer & J. T. Townsend (Eds.), *Oxford handbook of mathematical and computational psychology* (pp. 232–254). New York: Oxford University Press.

Kintsch, W. (2001). Predication. *Cognitive Science*, *25*(2), 173–202.

Kintsch, W., & Mangalath, P. (2011). The construction of meaning. *Topics in Cognitive Science*, *3*(2), 346–370.

Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 267–301). New York: Cambridge University Press.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin & Review*, *12*, 703–710.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, & C. Cortes (Eds.), *Advances in neural information processing systems* (pp. 2177–2185). Cambridge, MA: MIT Press Cambridge.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, andstimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775.

Murphy, G. L. (2016). Is there an exemplar theory of concepts? *Psychonomic Bulletin & Review*, *23*(4), 1035–1042.

Norman, G., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: The role of experience. *Medical Education*, *41*, 1140–1145.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2017). Tests of an Exemplar-Memory Model of Classification Learning in a High-Dimensional Natural-Science Category Domain. *Journal of Experimental Psychology: General*. http://psycnet.apa.org/doi/10.1037/xge0000369

Pickering, A. D. (1997). New approaches to the study of amnesic patients: What can a neurofunctional philosophy and neural network methods offer? *Memory*, *5*(1–2), 255–300.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3p1), 353–363.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407.

Reisinger, J., & Mooney, R. J. (2010). *Multi-prototype vector-space models of word meaning*. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, June 2010.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press

Rogers, T. T., & McClelland, J. L. (2011). 5 semantics without categorization. In E. M. Pothos & A. J. Wills (Eds.), *Formal Approaches in Categorization* (pp. 88–119). New York: Cambridge University Press.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4* (3), 328–350.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7* (4), 573–605.

Stanton, R. D., Nosofsky, R. M., & Zaki, S. R. (2002). Comparisons between exemplar similarity and mixed prototype models using a linearly separable category structure. *Memory & Cognition*, *30*(6), 934–944.

Thiessen, E. D. (2017). What's statistical about learning? Insights from modelling statistical learning as a set of memory processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160056.

Thiessen, E. D., & Pavlik, P. I. (2013). Iminerva: A mathematical model of distributional statistical learning. *Cognitive Science*, *37*(2), 310–343.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). New York, NY: Academic Press.

Valentin, V. V., Maddox, W. T., & Ashby, F. G. (2014). A computational model of the temporal dynamics of plasticity in procedural learning: sensitivity to feedback timing. *Frontiers in Psychology*, *5*, 1–9.

Vilnis, L., & McCallum, A. (2015). Word representations via gaussian embedding. arXiv preprint arXiv:1412.6623.

Weston, J., Chopra, S., & Bordes, A. (2015). Memory networks. arXiv preprint arXiv:1410.3916.