Topics in semantic representation

Thomas L. Griffiths
Department of Psychology
University of California, Berkeley

Mark Steyvers
Department of Cognitive Sciences
University of California, Irvine

Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

Accounts of language processing have suggested that it requires retrieving concepts from memory in response to an ongoing stream of information. This can be facilitated by inferring the gist of a sentence, conversation, or document, and using that gist to predict related concepts and disambiguate words. We analyze the abstract computational problem underlying the extraction and use of gist, formulating this problem as a rational statistical inference. This leads us to a novel approach to semantic representation in which word meanings are represented in terms of a set of probabilistic topics. The topic model performs well in predicting word association and the effects of semantic association and ambiguity on a variety of language processing and memory tasks. It also provides a foundation for developing more richly structured statistical models of language, as the generative process assumed in the topic model can easily be extended to incorporate other kinds of semantic and syntactic structure.

Many aspects of perception and cognition can be understood by considering the computational problem that is addressed by a particular human capacity (Andersion, 1990; Marr, 1982). Perceptual capacities such as identifying shape from shading (Freeman, 1994), motion perception (Weiss, Adelson, & Simoncelli, 2002), and sensorimotor integration (Wolpert, Ghahramani, & Jordan, 1995; Koerding & Wolpert, 2004) appear to closely approximate optimal statistical inferences.

This work was supported by a grant from the NTT Communication Sciences Laboratory and the DARPA CALO project. While completing this work, TLG was supported by a Stanford Graduate Fellowship and a grant from the National Science Foundation (BCS 0631518), and JBT by the Paul E. Newton chair. We thank Touchstone Applied Sciences, Tom Landauer, and Darrell Laham for making the TASA corpus available, and Steven Sloman and three anonymous reviewers for comments on the manuscript. A MATLAB toolbox containing code for simulating the various topic models described in this article is available at `http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm`.

Cognitive capacities such as memory and categorization can be seen as systems for efficiently making predictions about the properties of an organism's environment (e.g., Anderson, 1990). Solving problems of inference and prediction requires sensitivity to the statistics of the environment. Surprisingly subtle aspects of human vision can be explained in terms of the statistics of natural scenes (Geisler, Perry, Super, & Gallogly, 2001; Simoncelli & Olshausen, 2001), and human memory seems to be tuned to the probabilities with which particular events occur in the world (Anderson & Schooler, 1991). Sensitivity to relevant world statistics also seems to guide important classes of of cognitive judgments, such as inductive inferences about the properties of categories (Kemp, Perfors & Tenenbaum, 2004), predictions about the durations or magnitudes of events (Griffiths & Tenenbaum, 2006b), or inferences about hidden common causes from patterns of coincidence (Griffiths & Tenenbaum, 2006a).

In this paper, we examine how the statistics of one very important aspect of the environment – natural language – influence human memory. Our approach is motivated by an analysis of some of the computational problems addressed by semantic memory, in the spirit of Marr (1982) and Anderson (1990). Under many accounts of language processing, understanding sentences requires retrieving a variety of concepts from memory in response to an ongoing stream of information. One way to do this is to use the semantic context – the *gist* of a sentence, conversation, or document – to predict related concepts and disambiguate words (Ericsson & Kintsch, 1995; Kintsch, 1988; Potter, 1993). The retrieval of relevant information can be facilitated by predicting which concepts are likely to be relevant before they are needed. For example, if the word BANK appears in a sentence, it might become more likely that words like FEDERAL and RESERVE would also appear in that sentence, and this information could be used to initiate retrieval of the information related to these words. This prediction task is complicated by the fact that words have multiple senses or meanings: BANK should only influence the probabilities of FEDERAL and RESERVE if the gist of the sentence that it refers to a financial institution. If words like STREAM or MEADOW also appear in the sentence, then it is likely that BANK refers to the side of a river, and words like WOODS or FIELD should increase in probability.

The ability to extract gist has influences that reach beyond language processing, pervading even simple tasks such as memorizing lists of words. A number of studies have shown that when people try to remember a list of words that are semantically associated with a word that does not appear on the list, the associated word intrudes upon their memory (Deese, 1959; McEvoy, Nelson, & Komatsu, 1999; Roediger, Watson, McDermott, & Gallo, 2001). Results of this kind have led to the development of dual-route memory models, which suggest that people encode not just the verbatim content of a list of words, but also their gist (Brainerd, Reyna, & Mojardin, 1999; Brainerd, Wright, & Reyna, 2002; Mandler, 1980). These models leave open the question of how the memory system identifies this gist.

In this paper, we analyze the abstract computational problem of extracting and using the gist of a set of words, and examine how well different solutions to this problem correspond to human behavior. The key difference between these solutions is the way that they represent gist. In previous work, the extraction and use of gist has been modeled using associative semantic networks (e.g., Collins & Loftus, 1975) and semantic spaces (e.g., Landauer & Dumais, 1997; Lund & Burgess, 1996). Examples of these two representations are shown in Figure 1 (a) and (b), respectively. We take a step back from these specific proposals, and provide a more general formulation of the computational problem that these representations are used to solve. We express the problem as one of statistical inference: given some data – the set of words – inferring the latent structure from which

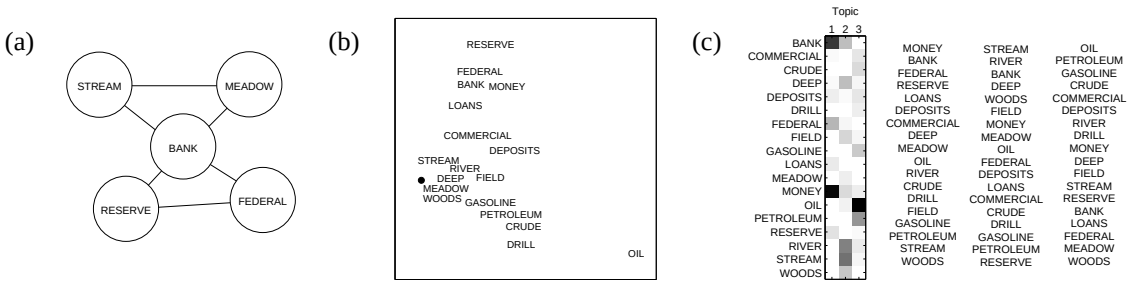

Figure 1. Approaches to semantic representation. (a) In a semantic network, words are represented as nodes, and edges indicate semantic relationships. (b) In a semantic space, words are represented as points, and proximity indicates semantic association. These are the first two dimensions of a solution produced by Latent Semantic Analysis (Landauer & Dumais, 1997). The black dot is the origin. (c) In the topic model, words are represented as belonging to a set of probabilistic topics. The matrix shown on the left indicates the probability of each word under each of three topics. The three columns on the right show the words that appear in those topics, ordered from highest to lowest probability.

it was generated. Stating the problem in these terms makes it possible to explore forms of semantic representation that go beyond networks and spaces.

Identifying the statistical problem underlying the extraction and use of gist makes it possible to use any form of semantic representation: all that needs to be specified is a probabilistic process by which a set of words are generated using that representation of their gist. In machine learning and statistics, such a probabilistic process is called a *generative model*. Most computational approaches to natural language have tended to focus exclusively on either structured representations (e.g., Chomsky, 1965; Pinker, 1999) or statistical learning (e.g., Elman, 1990; Plunkett & Marchman, 1993; Rumelhart & McClelland, 1986). Generative models provide a way to combine the strengths of these two traditions, making it possible to use statistical methods to learn structured representations. As a consequence, generative models have recently become popular in both computational linguistics (e.g., Charniak, 1993; Jurafsky & Martin, 2000; Manning & Shütze, 1999) and psycholinguistics (e.g., Baldewein & Keller, 2004; Jurafsky, 1996), although this work has tended to emphasize syntactic structure over semantics.

The combination of structured representations with statistical inference makes generative models the perfect tool for evaluating novel approaches to semantic representation. We use our formal framework to explore the idea that the gist of a set of words can be represented as a probability distribution over a set of topics. Each topic is a probability distribution over words, and the content of the topic is reflected in the words to which it assigns high probability. For example, high probabilities for WOODS and STREAM would suggest a topic refers to the countryside, while high probabilities for FEDERAL and RESERVE would suggest a topic refers to finance. A schematic illustration of this form of representation appears in Figure 1 (c). Following work in the information retrieval literature (Blei, Ng, & Jordan, 2003), we use a simple generative model that defines a probability distribution over a set of words, such as a list or a document, given a probability distribution over topics. Using methods from Bayesian statistics, a set of topics can be learned automatically from a collection of documents, as a computational analog of how human learners might form semantic representations through their linguistic experience (Griffiths & Steyvers, 2002, 2003, 2004).

The topic model provides a starting point for an investigation of new forms of semantic representation. Representing words using topics has an intuitive correspondence to feature-based models of similarity. Words that receive high probability under the same topics will tend to be highly predictive of one another, just as stimuli that share many features will be highly similar. We show that this intuitive correspondence is supported by a formal correspondence between the topic model and Tversky's (1977) feature-based approach to modeling similarity. Since the topic model uses exactly the same input as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), a leading model of the acquisition of semantic knowledge in which the association between words depends on the distance between them in a semantic space, we can compare these two models as a means of examining the implications of different kinds of semantic representation, just as featural and spatial representations have been compared as models of human similarity judgments (Tversky, 1977; Tversky & Gati, 1982; Tversky & Hutchinson, 1986). Furthermore, the topic model can easily be extended to capture other kinds of latent linguistic structure. Introducing new elements into a generative model is straightforward, and by adding components to the model that can capture richer semantic structure or rudimentary syntax we can begin to develop more powerful statistical models of language.

The plan of the paper is as follows. First, we provide a more detailed specification of the kind of semantic information we aim to capture in our models, and summarize the ways in which this has been done in previous work. We then analyze the abstract computational problem of extracting and using gist, formulating this problem as one of statistical inference and introducing the topic model as one means of solving this computational problem. The body of the paper is concerned with assessing how well the representation recovered by the topic model corresponds with human semantic memory. In an analysis inspired by Tversky's (1977) critique of spatial measures of similarity, we show that several aspects of word association that can be explained by the topic model are problematic for LSA. We then compare the performance of the two models in a variety of other tasks tapping semantic representation, and outline some of the way in which the topic model can be extended.

Approaches to semantic representation

Semantic representation is one of the most formidable topics in cognitive psychology. The field is fraught with murky and potentially never-ending debates; it is hard to imagine that one could give a complete theory of semantic representation outside of a complete theory of cognition in general. Consequently, formal approaches to modeling semantic representation have focused on various tractable aspects of semantic knowledge. Before presenting our approach we must clarify where its focus lies.

Semantic knowledge can be thought of as knowledge about relations among several types of elements, including words, concepts, and percepts. Some relations that have been studied include the following:

Word-concept relations. Knowledge that the word DOG refers to the concept *dog*, the word ANIMAL refers to the concept *animal*, or that the word TOASTER refers to the concept *toaster*.

Concept-concept relations. Knowledge that *dogs* are a kind of *animal*, that *dogs* have *tails* and can *bark*, or that *animals* have *bodies* and can *move*.

Concept-percept or concept-action relations. Knowledge about what *dogs* look like, how a *dog* can be distinguished from a *cat*, how to pet a *dog* or operate a *toaster*.

Word-word relations. Knowledge that the word DOG tends to be associated with or co-occur with words such as TAIL, BONE or CAT, or the word TOASTER tends to be associated with KITCHEN, OVEN, or BREAD.

These different aspects of semantic knowledge are not necessarily independent. For instance, the word CAT may be associated with the word DOG because CAT refers to *cats*, DOG refers to *dogs*, and *cats* and *dogs* are both common kinds of *animals*. Yet different aspects of semantic knowledge can influence behavior in different ways and seem to be best captured by different kinds of formal representations. As a result, different approaches to modeling semantic knowledge tend to focus on different aspects of this knowledge, depending on what fits most naturally with the representational system they adopt, and there are corresponding differences in the behavioral phenomena they emphasize. Computational models also differ in the extent to which their semantic representations can be learned automatically from some naturally occuring data or must be hand-wired by the modeler. Although many different modeling approaches can be imagined within this broad landscape, there are two prominent traditions.

One tradition emphasizes abstract conceptual structure, focusing on relations between concepts and relations between concepts and percepts or actions. This knowledge is traditionally represented in terms of systems of abstract propositions, such as (is-a *canary bird*), (has *bird wings*), and so on (Collins & Quillian, 1969). Models in this tradition have focused on explaining phenomena such as the development of conceptual hierarchies that support propositional knowledge (e.g., Keil, 1979), reaction time to verify conceptual propositions in normal adults (Collins & Quillian, 1969), and the decay of propositional knowledge with aging or brain damage (e.g., Warrington, 1975). This approach does not worry much about the mappings between words and concepts, or associative relations between words; in practice, the distinction between words and concepts is typically collapsed. Actual language use is addressed only indirectly: the relevant experiments are often conducted with linguistic stimuli and responses, but the primary interest is not in the relation between language use and conceptual structure. Representations of abstract semantic knowledge of this kind have traditionally been hand-crafted by modelers (Collins & Quillian, 1969), in part because it is not clear how they could be learned automatically. Recently there has been some progress in learning distributed representations of conceptual relations (Rogers & McClelland, 2004), although the input to these learning models is still quite idealized, in the form of hand-coded databases of simple propositions. Learning large-scale representations of abstract conceptual relations from naturally occuring data remains an unsolved problem.

A second tradition of studying semantic representation has focused more on the structure of associative relations between words in natural language use, and relations between words and concepts, along with the contextual dependence of these relations. For instance, when one hears the word BIRD, it becomes more likely that one will also hear words like SING, FLY, or NEST in the same context – but perhaps less so if the context also contains the words THANKSGIVING, TURKEY, and DINNER. These expectations reflect the fact that BIRD has multiple senses, or multiple concepts it can refer to, including both a taxonomic category and a food category. The semantic phenomena studied in this tradition may appear to be somewhat superficial, in that they typically do not tap deep conceptual understanding. The data tend to be tied more directly to language use and the memory systems that support online linguistic processing, such as word association norms

(e.g., Nelson, McEvoy, & Schreiber, 1998), word reading times in sentence processing (e.g., Sereno, Pacht, & Rayner, 1992), semantic priming (e.g., Till, Mross, & Kintsch, 1988), and effects of semantic context in free recall (e.g., Roediger & McDermott, 1995). Compared to approaches focusing on deeper conceptual relations, classic models of semantic association tend to invoke much simpler semantic representations, such as semantic spaces or holistic spreading-activation networks (e.g., Deese, 1959; Collins & Loftus, 1975). This simplicity has its advantages: there has recently been considerable success in learning the structure of such models from large-scale linguistic corpora (e.g., Landauer & Dumais, 1997; Lund & Burgess, 1996).

We recognize the importance of both these traditions in studying semantic knowledge. They have complementary strengths and weaknesess, and ultimately ideas from both are likely to be important. Our work here is more clearly in the second tradition, with its emphasis on relatively light representations that can be learned from large text corpora, and on explaining the structure of word-word and word-concept associations, rooted in the contexts of actual language use. While the interpretation of sentences requires semantic knowledge that goes beyond these contextual associative relationships, many theories still identify this level of knowledge as playing an important role in the early stages of language processing (Ericsson & Kintsch, 1995; Kintsch, 1988; Potter, 1993). Specifically, it supports solutions to three core computational problems:

Prediction	Predict the next word or concept, facilitating retrieval
Disambiguation	Identify the senses or meanings of words
Gist extraction	Pick out the gist of a set of words

Our goal is to understand how contextual semantic association is represented, used, and acquired. We will argue that considering relations between latent semantic topics and observable word forms provides a way to capture many aspects of this level of knowledge: it provides principled and powerful solutions to these three core tasks and it is also easily learnable from natural linguistic experience. Before introducing this modeling framework, we will summarize the two dominant approaches to the representation of semantic association, semantic networks and semantic spaces, establishing the background to the problems we consider.

Semantic networks

In an associative semantic network, such as that shown in Figure 1 (a), a set of words or concepts are represented as nodes connected by edges that indicate pairwise associations (e.g., Collins & Loftus, 1975). Seeing a word activates its node, and activation spreads through the network, activating nodes that are nearby. Semantic networks provide an intuitive framework for expressing the semantic relationships between words. They also provide simple solutions to the three problems for which contextual knowledge might be used. Treating those problems in the reverse of the order identified above, gist extraction simply consists of activating each word that occurs in a given context, and allowing that activation to spread through the network. The gist is represented by the pattern of node activities. If different meanings of words are represented as different nodes, then disambiguation can be done by comparing the activation of those nodes. Finally, the words that one might expect to see next in that context will be the words that have high activations as a result of this process.

Most semantic networks that are used as components of cognitive models are considerably more complex than the example shown in Figure 1 (a), allowing multiple different kinds of nodes and connections (e.g., Anderson, 1983; Norman, Rumelhart, & the LNR Research Group, 1975).

In addition to "excitatory" connections, in which activation of one node increases activation of another, some semantic networks feature "inhibitory" connections, allowing activation of one node to decrease activation of another. The need for inhibitory connections is indicated by empirical results in the literature on priming. A simple network without inhibitory connections can explain why priming might facilitate lexical decision, making it easier to recognize that a target is an English word. For example, a word like NURSE primes the word DOCTOR because it activates concepts that are closely related to DOCTOR, and the spread of activation ultimately activates doctor. However, not all priming effects are of this form. For example, Neely (1976) showed that priming with irrelevant cues could have an inhibitory effect on lexical decision. To use an example from Markman (1998), priming with HOCKEY could produce a slower reaction time for DOCTOR than presenting a completely neutral prime. Effects like these suggest that we need to incorporate inhibitory links between words. Interestingly, it would seem that a great many such links would be required, because there is no obvious special relationship between HOCKEY and DOCTOR: the two words just seem unrelated. Thus, inhibitory links would seem to be needed between all pairs of unrelated words in order to explain inhibitory priming.

Semantic spaces

An alternative to semantic networks is the idea that the meaning of words can be captured using a spatial representation. In a semantic space, such as that shown in Figure 1 (b), words are nearby if they are similar in meaning. This idea appears in early work exploring the use of statistical methods to extract representations of the meaning of words from human judgments (Deese, 1959; Fillenbaum & Rapoport, 1971). Recent research has pushed this idea in two directions. First, connectionist models using "distributed representations" for words – which are commonly interpreted as a form of spatial representation – have been used to predict behavior on a variety of linguistic tasks (e.g., Kawamoto, 1993; Plaut, 1997; Rodd, Gaskell, & Marslen-Wilson, 2004). These models perform relatively complex computations on the underlying representations and allow words to be represented as multiple points in space, but are typically trained on artificially generated data. A second thrust of recent research has been exploring methods for extracting semantic spaces directly from real linguistic corpora (Landauer & Dumais, 1997; Lund & Burgess, 1996). These methods are based upon comparatively simple models – for example, they assume each word is represented as only a single point – but provide a direct means of investigating the influence of the statistics of language on semantic representation.

Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) is one of the most prominent methods for extracting a spatial representation for words from a multi-document corpus of text. The input to LSA is a word-document co-occurrence matrix, such as that shown in Figure 2. In a word-document co-occurrence matrix, each row represents a word, each column represents a document, and the entries indicate the frequency with which that word occurred in that document. The matrix shown in Figure 2 is a portion of the full co-occurrence matrix for the TASA corpus (Landauer & Dumais, 1997), a collection of passages excerpted from educational texts used in curricula from the first year of school to the first year of college.

The output from LSA is a spatial representation for words and documents. After applying various transformations to the entries in a word-document co-occurrence matrix (one standard set of transformations is described in Griffiths & Steyvers, 2003), singular value decomposition is used to factorize this matrix into three smaller matrices, U, D, and V, as shown in Figure 3 (a). Each of these matrices has a different interpretation. The U matrix provides an orthonormal basis for

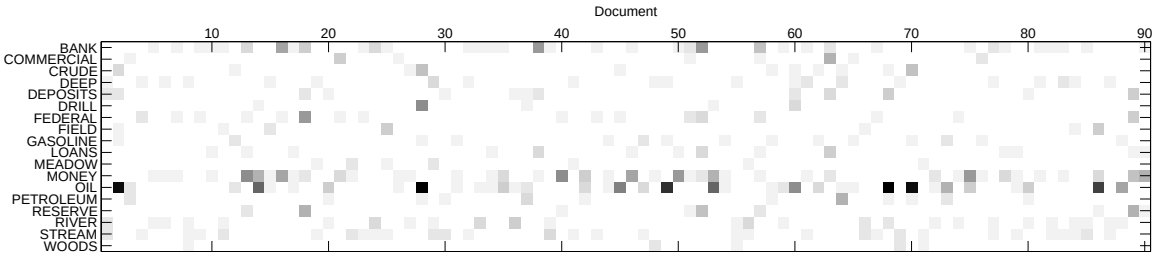

Figure 2. A word-document co-occurrence matrix, indicating the frequencies of 18 words across 90 documents extracted from the TASA corpus. A total of 30 documents use the word MONEY, 30 use the word OIL, and 30 use the word RIVER. Each row corresponds to a word in the vocabulary, and each column to a document in the corpus. Grayscale indicates the frequency with which the 731 tokens of those words appeared in the 90 documents, with black being the highest frequency and white being zero.

a space in which each word is a point. The D matrix, which is diagonal, is a set of weights for the dimensions of this space. The V matrix provides an orthonormal basis for a space in which each document is a point. An approximation to the original matrix of transformed counts can be obtained by remultiplying these matrices, but choosing to use only the initial portions of each matrix, corresponding to the use of a lower-dimensional spatial representation.

In psychological applications of LSA, the critical result of this procedure is the first matrix, U, which provides a spatial representation for words. Figure 1 (b) shows the first two dimensions of U for the word-document co-occurence matrix shown in Figure 2. The results shown in the figure demonstrate that LSA identifies some appropriate clusters of words. For example, OIL, PETROLEUM and CRUDE are close together, as are FEDERAL, MONEY, and RESERVE. The word DEPOSITS lies between the two clusters, reflecting the fact that it can appear in either context.

The cosine of the angle between the vectors corresponding to words in the semantic space defined by U has proven to be an effective measure of the semantic association between those words (Landauer & Dumais, 1997). The cosine of the angle between two vectors w_1 and w_2 (both rows of U, converted to column vectors) is

$$\cos(w_1, w_2) = \frac{w_1^T w_2}{||w_1|| ||w_2||}, \tag{1}$$

where $w_1^T w_2$ is the inner product of the vectors w_1 and w_2, and $||w||$ denotes the norm, $\sqrt{w^T w}$. Performance in predicting human judgments is typically better when using only the first few hundred derived dimensions, since reducing the dimensionality of the representation can decrease the effects of statistical noise and emphasize the latent correlations among words (Landauer & Dumais, 1997).

Latent Semantic Analysis provides a simple procedure for extracting a spatial representation of the associations between words from a word-document co-occurrence matrix. The gist of a set of words is represented by the average of the vectors associated with those words. Applications of LSA often evaluate the similarity between two documents by computing the cosine between the average word vectors for those documents (Landauer & Dumais, 1997; Rehder, Schreiner, Wolfe, Laham, Landauer, & Kintsch, 1998; Wolfe, Schreiner, Rehder, Laham, Foltz, Kintsch, & Landauer, 1998). This representation of the gist of a set of words can be used to address the prediction problem: we should predict that words with vectors close to the gist vector are likely to occur in the same context. However, the representation of words as points in an undifferentiated Euclidean space makes it

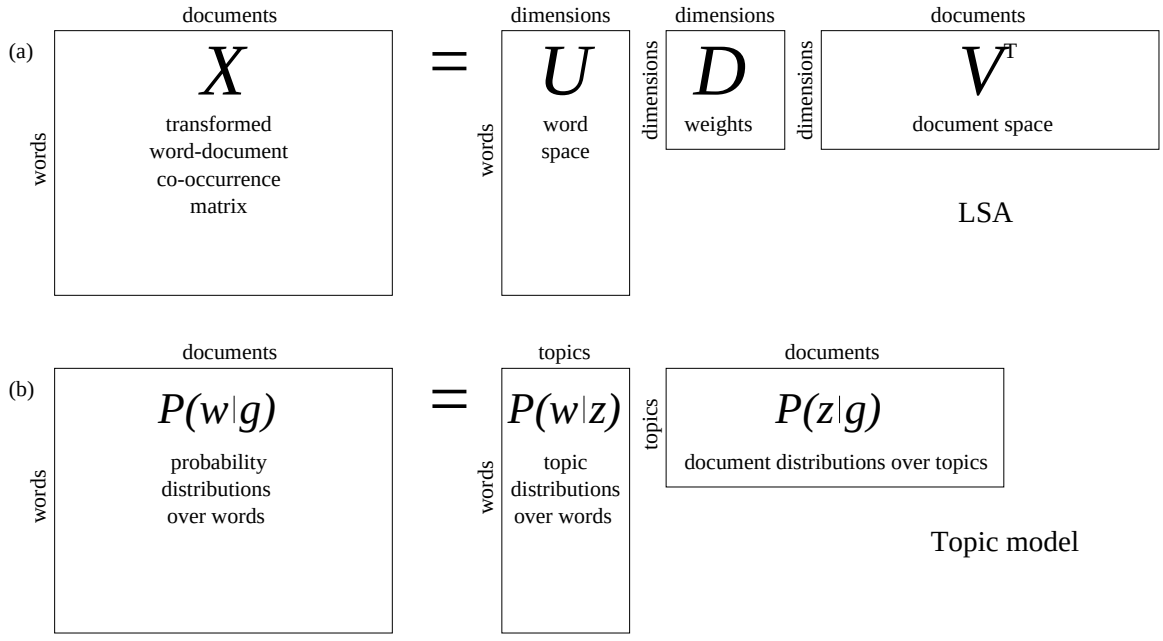

Figure 3. (a) Latent Semantic Analysis (LSA) performs dimensionality reduction using the singular value decomposition. The transformed word-document co-occurrence matrix, X, is factorized into three smaller matrices, U, D, and V. U provides an orthonormal basis for a spatial representation of words, D weights those dimensions, and V provides an orthonormal basis for a spatial representation of documents. (b) The topic model performs dimensionality reduction using statistical inference. The probability distribution over words for each document in the corpus conditioned upon its gist, $P(w|g)$, is approximated by a weighted sum over a set of probabilistic topics, represented with probability distributions over words $P(w|z)$, where the weights for each document are probability distributions over topics, $P(z|g)$, determined by the gist of the document, g.

difficult for LSA to solve the disambiguation problem. The key issue is that this relatively unstructured representation does not explicitly identify the different senses of words. While DEPOSITS lies between words having to do with finance and words having to do with oil, the fact that this word has multiple senses is not encoded in the representation.

Extracting and using gist as statistical problems

Semantic networks and semantic spaces are both proposals for a form of semantic representation that can guide linguistic processing. We now take a step back from these specific proposals, and consider the abstract computational problem that they are intended to solve, in the spirit of Marr's (1982) notion of the computational level, and Anderson's (1990) rational analysis. Our aim is to clarify the goals of the computation and to identify the logic by which these goals can be achieved, so that this logic can be used as the basis for exploring other approaches to semantic representation.

Assume we have seen a sequence of words $\mathbf{w} = (w_1, w_2, \ldots, w_n)$. These n words manifest some latent semantic structure $\boldsymbol{\ell}$. We will assume that $\boldsymbol{\ell}$ consists of the gist of that sequence of words g, and the sense or meaning of each word, $\mathbf{z} = (z_1, z_2, \ldots, z_n)$, so $\boldsymbol{\ell} = (g, \mathbf{z})$. We can now formalize the three problems identified in the previous section:

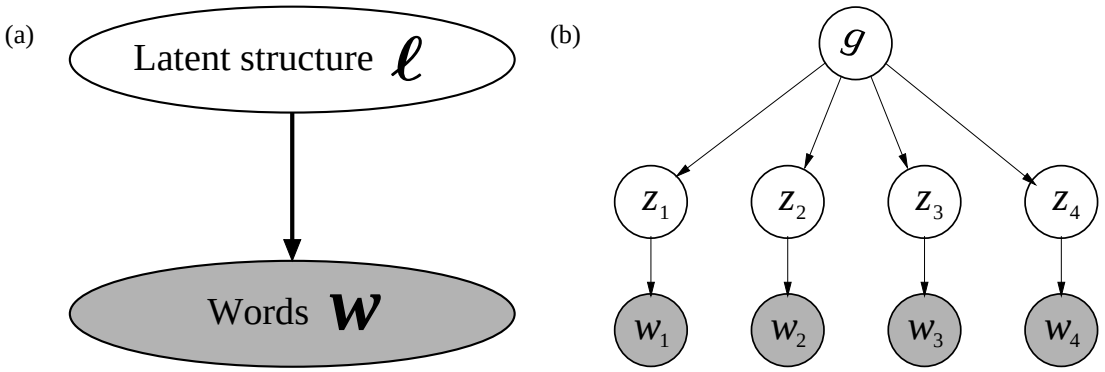

Figure 4. Generative models for language. (a) A schematic representation of generative models for language. Latent structure ℓ generates words $\mathbf{w}$. This generative process defines a probability distribution over ℓ, $P(\ell)$, and $\mathbf{w}$ given ℓ, $P(\mathbf{w}|\ell)$. Applying Bayes' rule with these distributions makes it possible to invert the generative process, inferring ℓ from $\mathbf{w}$. (b) Latent Dirichlet Allocation (Blei et al., 2003), a topic model. A document is generated by choosing a distribution over topics that reflects the gist of the document, g, choosing a topic z_i for each potential word from a distribution determined by g, and then choosing the actual word w_i from a distribution determined by z_i.

Prediction	Predict w_{n+1} from $\mathbf{w}$
Disambiguation	Infer $\mathbf{z}$ from $\mathbf{w}$
Gist extraction	Infer g from $\mathbf{w}$

Each of these problems can be formulated as statistical problems. The prediction problem requires computing the conditional probability of w_{n+1} given $\mathbf{w}$, $P(w_{n+1}|\mathbf{w})$. The disambiguation problem requires computing the conditional probability of $\mathbf{z}$ given $\mathbf{w}$, $P(\mathbf{z}|\mathbf{w})$. The gist extraction problem requires computing the probability of g given $\mathbf{w}$, $P(g|\mathbf{w})$.

All of the probabilities needed to solve the problems of prediction, disambiguation, and gist extraction can be computed from a single joint distribution over words and latent structures, $P(\mathbf{w}, \ell)$. The problems of prediction, disambiguation, and gist extraction can thus be solved by learning the joint probabilities of words and latent structures. This can be done using a generative model for language. Generative models are widely used in machine learning and statistics as a means of learning structured probability distributions. A generative model specifies a hypothetical causal process by which data are generated, breaking this process down into probabilistic steps. Critically, this procedure can involve unobserved variables, corresponding to latent structure that plays a role in generating the observed data. Statistical inference can be used to identify the latent structure most likely to have been responsible for a set of observations.

A schematic generative model for language is shown in Figure 4 (a). In this model, latent structure ℓ generates an observed sequence of words $\mathbf{w} = (w_1, \ldots, w_n)$. This relationship is illustrated using *graphical model* notation (e.g., Jordan, 1998; Pearl, 1988). Graphical models provide an efficient and intuitive method of illustrating structured probability distributions. In a graphical model, a distribution is associated with a graph in which nodes are random variables and edges indicate dependence. Unlike artificial neural networks, in which a node typically indicates a single unidimensional variable, the variables associated with nodes can be arbitrarily complex. ℓ can be any kind of latent structure, and $\mathbf{w}$ represents a set of n words.

The graphical model shown in Figure 4 (a) is a *directed* graphical model, with arrows indi-

cating the direction of the relationship among the variables. The result is a directed graph, in which "parent" nodes have arrows to their "children". In a generative model, the direction of these arrows specifies the direction of the causal process by which data are generated: a value is chosen for each variable by sampling from a distribution that conditions on the parents of that variable in the graph. The graphical model shown in the figure indicates that words are generated by first sampling a latent structure, $\boldsymbol{\ell}$, from a distribution over latent structures, $P(\boldsymbol{\ell})$, and then sampling a sequence of words, $\mathbf{w}$, conditioned on that structure from a distribution $P(\mathbf{w}|\boldsymbol{\ell})$.

The process of choosing each variable from a distribution conditioned on its parents defines a joint distribution over observed data and latent structures. In the generative model shown in Figure 4 (a), this joint distribution is

$$P(\mathbf{w}, \boldsymbol{\ell}) = P(\mathbf{w}|\boldsymbol{\ell})P(\boldsymbol{\ell}).$$

With an appropriate choice of $\boldsymbol{\ell}$, this joint distribution can be used to solve the problems of prediction, disambiguation, and gist extraction identified above. In particular, the probability of the latent structure $\boldsymbol{\ell}$ given the sequence of words $\mathbf{w}$ can be computed by applying Bayes' rule:

$$P(\boldsymbol{\ell}|\mathbf{w}) = \frac{P(\mathbf{w}|\boldsymbol{\ell})P(\boldsymbol{\ell})}{P(\mathbf{w})} \tag{2}$$

where

$$P(\mathbf{w}) = \sum_{\boldsymbol{\ell}} P(\mathbf{w}|\boldsymbol{\ell})P(\boldsymbol{\ell}).$$

This Bayesian inference involves computing a probability that goes against the direction of the arrows in the graphical model, inverting the generative process.

Equation 2 provides the foundation for solving the problems of prediction, disambiguation, and gist extraction. The probability needed for prediction, $P(w_{n+1}|\mathbf{w})$, can be written as

$$P(w_{n+1}|\mathbf{w}) = \sum_{\boldsymbol{\ell}} P(w_{n+1}|\boldsymbol{\ell}, \mathbf{w})P(\boldsymbol{\ell}|\mathbf{w}), \tag{3}$$

where $P(w_{n+1}|\boldsymbol{\ell})$ is specified by the generative process. Distributions over the senses of words, $\mathbf{z}$, and their gist, g, can be computed by summing out the irrelevant aspect of $\boldsymbol{\ell}$,

$$P(\mathbf{z}|\mathbf{w}) = \sum_{g} P(\boldsymbol{\ell}|\mathbf{w}) \tag{4}$$

$$P(g|\mathbf{w}) = \sum_{\mathbf{z}} P(\boldsymbol{\ell}|\mathbf{w}), \tag{5}$$

where we assume that the gist of a set of words takes on a discrete set of values – if it is continuous, then Equation 5 requires an integral rather than a sum.

This abstract schema gives a general form common to all generative models for language. Specific models differ in the latent structure $\boldsymbol{\ell}$ that they assume, the process by which this latent structure is generated (which defines $P(\boldsymbol{\ell})$), and the process by which words are generated from this latent structure (which defines $P(\mathbf{w}|\boldsymbol{\ell})$). Most generative models that have been applied to language focus on latent syntactic structure (e.g., Charniak, 1993; Jurafsky & Martin, 2000; Manning & Shütze, 1999). In the next section, we describe a generative model that represents the latent semantic structure that underlies a set of words.

Representing gist with topics

A topic model is a generative model that assumes a latent structure $\ell = (g, \mathbf{z})$, representing the gist of a set of words, g, as a distribution over T topics, and the sense or meaning used for the ith word, z_i, as an assignment of that word to one of these topics.[1] Each topic is a probability distribution over words. A document – a set of words – is generated by choosing the distribution over topics reflecting its gist, using this distribution to choose a topic z_i for each word w_i, and then generating the word itself from the distribution over words associated with that topic. Given the gist of the document in which it is contained, this generative process defines the probability of the ith word to be

$$P(w_i|g) = \sum_{z_i=1}^{T} P(w_i|z_i)P(z_i|g), \tag{6}$$

in which the topics, specified by $P(w|z)$, are mixed together with weights given by $P(z|g)$, which vary across documents.[2] The dependency structure among variables in this generative model is shown in Figure 4 (b).

Intuitively, $P(w|z)$ indicates which words are important to a topic, while $P(z|g)$ is the prevalence of those topics in a document. For example, if we lived in a world where people only wrote about finance, the English countryside, and oil mining, then we could model all documents with the three topics shown in Figure 1 (c). The content of the three topics is reflected in $P(w|z)$: the finance topic gives high probability to words like RESERVE and FEDERAL, the countryside topic gives high probability to words like STREAM and MEADOW, and the oil topic gives high probability to words like PETROLEUM and GASOLINE. The gist of a document, g, indicates whether a particular document concerns finance, the countryside, oil mining, or financing an oil refinery in Leicestershire, by determining the distribution over topics, $P(z|g)$.

Equation 6 gives the probability of a word conditioned on the gist of a document. We can define a generative model for a collection of documents by specifying how the gist of each document is chosen. Since the gist is a distribution over topics, this requires using a distribution over multinomial distributions. The idea of representing documents as mixtures of probabilistic topics has been used in a number of applications in information retrieval and statistical natural language processing, with different models making different assumptions about the origins of the distribution over topics (e.g., Bigi, De Mori, El Beze, & Spriet, 1997; Blei et al., 2003; Hofmann, 1999; Iyer & Ostendorf, 1996; Ueda & Saito, 2003). We will use a generative model introduced by Blei et al. (2003) called Latent Dirichlet Allocation. In this model, the multinomial distribution representing the gist is drawn from a from a Dirichlet distribution, a standard probability distribution over multinomials (e.g., Gelman, Carlin, Stern, & Rubin, 1995).

Having defined a generative model for a corpus based upon some parameters, it is possible to use statistical methods to infer the parameters from the corpus. In our case, this means finding a set

[1] This formulation of the model makes the assumption that each topic captures a different sense or meaning of a word. This need not be the case – there may be a many-to-one relationship between topics and the senses or meanings in which words are used. However, the topic assignment still communicates information that can be used in disambiguation and prediction in the way that the sense or meaning must be used. Henceforth, we will focus on the use of z_i to indicate a topic assignment, rather than a sense or meaning for a particular word.

[2] We have suppressed the dependence of the probabilities discussed in this section on the parameters specifying $P(w|z)$ and $P(z|g)$, assuming that these parameters are known. A more rigorous treatment of the computation of these probabilities is given in Appendix A.

of topics such that each document can be expressed as a mixture of those topics. An algorithm for extracting a set of topics is described in Appendix A, and a more detailed description and application of this algorithm can be found in Griffiths and Steyvers (2004). This algorithm takes as input a word-document co-occurrence matrix. The output is a set of topics, each being a probability distribution over words. The topics shown in Figure 1 (c) are actually the output of this algorithm when applied to the word-document co-occurrence matrix shown in Figure 2. These results illustrate how well the topic model handles words with multiple meanings or senses: FIELD appears in both the oil and countryside topics, BANK appears in both finance and countryside, and DEPOSITS appears in both oil and finance. This is a key advantage of the topic model: by assuming a more structured representation, in which words are assumed to belong to topics, the different meanings or senses of ambiguous words can be differentiated.

Prediction, disambiguation, and gist extraction

The topic model provides a direct solution to the problems of prediction, disambiguation, and gist extraction identified in the previous section. The details of these computations are presented in Appendix A. To illustrate how these problems are solved by the model, we will consider a simplified case where all words in a sentence are assumed to have the same topic. In this case g is a distribution that puts all of its probability on a single topic, z, and $z_i = z$ for all i. This "single topic" assumption makes the mathematics straightforward, and is a reasonable working assumption in many of the settings we explore.[3]

Under the single topic assumption, disambiguation and gist extraction become equivalent: the senses and the gist of a set of words are both expressed in the single topic, z, that was responsible for generating words $\mathbf{w} = \{w_1, w_2, \ldots, w_n\}$. Applying Bayes' rule, we have

$$\begin{aligned} P(z|\mathbf{w}) &= \frac{P(\mathbf{w}|z)P(z)}{P(\mathbf{w})} \\ &= \frac{\prod_{i=1}^{n} P(w_i|z)P(z)}{\sum_z \prod_{i=1}^{n} P(w_i|z)P(z)}, \end{aligned} \tag{7}$$

where we have used the fact that the w_i are independent given z. If we assume a uniform prior over topics, $P(z) = \frac{1}{T}$, the distribution over topics depends only on the product of the probabilities of each of the w_i under each topic z. The product acts like a logical "and": a topic will only be likely if it gives reasonably high probability to all of the words. Figure 5 shows how this functions to disambiguate words, using the topics from Figure 1. On seeing the word BANK, both the finance and the countryside topics have high probability. Seeing STREAM quickly swings the probability in favor of the bucolic interpretation.

Solving the disambiguation problem is the first step in solving the prediction problem. Incorporating the assumption that words are independent given their topics into Equation 3, we have

$$P(w_{n+1}|\mathbf{w}) = \sum_z P(w_{n+1}|z)P(z|\mathbf{w}). \tag{8}$$

[3] It is also possible to define a generative model that makes this assumption directly, having just one topic per sentence, and to use techniques like those described in Appendix A to identify topics using this model. We did not use this model because it uses additional information about the structure of the documents, making it harder to compare against alternative approaches such as Latent Semantic Analysis (Landauer & Dumais, 1997). The single topic assumption can also be derived as the consequence of having a hyperparameter α favoring choices of $\mathbf{z}$ that employ few topics: the single topic assumption is produced by allowing α to approach 0.

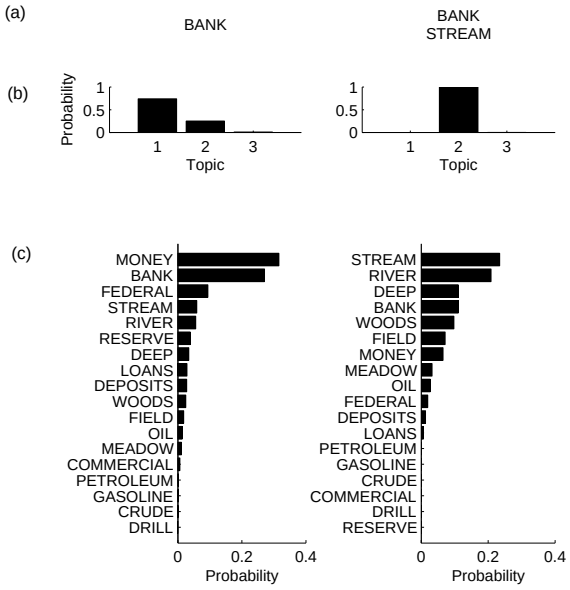

Figure 5. Prediction and disambiguation. (a) Words observed in a sentence, $\mathbf{w}$. (b) The distribution over topics conditioned on those words, $P(z|\mathbf{w})$. (c) The predicted distribution over words resulting from summing over this distribution over topics, $P(w_{n+1}|\mathbf{w}) = \sum_z P(w_{n+1}|z)P(z|\mathbf{w})$. On seeing BANK, the model is unsure whether the sentence concerns finance or the countryside. Subsequently seeing STREAM results in a strong conviction that BANK does not refer to a financial institution.

The predicted distribution over words is thus a mixture of topics, with each topic being weighted by the distribution computed in Equation 7. This is illustrated in Figure 5: on seeing BANK, the predicted distribution over words is a mixture of the finance and countryside topics, but STREAM moves this distribution towards the countryside topic.

Topics and semantic networks

The topic model provides a clear way of thinking about how and why "activation" might spread through a semantic network, and can also explain inhibitory priming effects. The standard conception of a semantic network is a graph with edges between word nodes, as shown in Figure 6 (a). Such a graph is *unipartite*: there is only one type of node, and those nodes can be interconnected freely. In contrast, *bipartite* graphs consist of nodes of two types, and only nodes of different types can be connected. We can form a bipartite semantic network by introducing a second class of nodes that mediate the connections between words. One way to think about the representation of the meanings of words provided by the topic model is in terms of the bipartite semantic network shown in Figure 6 (b), where the second class of nodes are the topics.

In any context, there is uncertainty about which topics are relevant to that context. On seeing a word, the probability distribution over topics moves to favor the topics associated with that word: $P(z|\mathbf{w})$ moves away from uniformity. This increase in the probability of those topics is intuitively similar to the idea that activation spreads from the words to the topics that are connected with them. Following Equation 8, the words associated with those topics also receive higher probability.

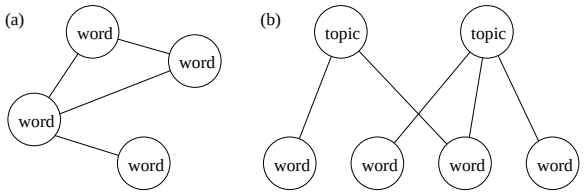

Figure 6. Semantic networks. (a) In a unipartite network, there is only one class of nodes. In this case, all nodes represent words. (b) In a bipartite network, there are two classes, and connections only exist between nodes of different classes. In this case, one class of nodes represents words and the other class represents topics.

This dispersion of probability throughout the network is again reminiscent of spreading activation. However, there is an important difference between spreading activation and probabilistic inference: the probability distribution over topics, $P(z|\mathbf{w})$ is constrained to sum to one. This means that as the probability of one topic increases, the probability of another topic decreases.

The constraint that the probability distribution over topics sums to one is sufficient to produce the phenomenon of inhibitory priming discussed above. Inhibitory priming occurs as a necessary consequence of excitatory priming: when the probability of one topic increases, the probability of another topic decreases. Consequently, it is possible for one word to decrease the predicted probability with which another word will occur in a particular context. For example, according to the topic model, the probability of the word DOCTOR is 0.000334. Under the single topic assumption, the probability of the word DOCTOR conditioned on the word NURSE is 0.0071, an instance of excitatory priming. However, the probability of DOCTOR drops to 0.000081 when conditioned on HOCKEY. The word HOCKEY suggests that the topic concerns sports, and consequently topics that give DOCTOR high probability have lower weight in making predictions. By incorporating the constraint that probabilities sum to one, generative models are able to capture both the excitatory and the inhibitory influence of information without requiring the introduction of large numbers of inhibitory links between unrelated words.

Topics and semantic spaces

Our claim that models that can accurately predict which words are likely to arise in a given context can provide clues about human language processing is shared with the spirit of many connectionist models (e.g., Elman, 1990). However, the strongest parallels between our approach and work being done on spatial representations of semantics are perhaps those that exist between the topic model and Latent Semantic Analysis (Landauer & Dumais, 1997). Indeed, the probabilistic topic model developed by Hofmann (1999) was motivated by the success of LSA, and provided the inspiration for the model introduced by Blei et al. (2003) that we use here. Both LSA and the topic model take a word-document co-occurrence matrix as input. Both LSA and the topic model provide a representation of the gist of a document, either as a point in space or a distribution over topics. And both LSA and the topic model can be viewed as a form of "dimensionality reduction", attempting to find a lower-dimensional representation of the structure expressed in a collection of documents. In the topic model, this dimensionality reduction consists of trying to express the large number of probability distributions over words provided by the different documents in terms of a small number of topics, as illustrated in Figure 3 (b).

However, there are two important differences between LSA and the topic model. The major difference is that LSA is not a generative model. It does not identify a hypothetical causal process responsible for generating documents, and the role of the meanings of words in this process. As a consequence, it is difficult to extend LSA to incorporate different kinds of semantic structure, or to recognize the syntactic roles that words play in a document. This leads to the second difference between LSA and the topic model: the nature of the representation. Latent Semantic Analysis is based upon the singular value decomposition, a method from linear algebra that can only yield a representation of the meanings of words as points in an undifferentiated Euclidean space. In contrast, the statistical inference techniques used with generative models are flexible, and make it possible to use structured representations. The topic model provides a simple structured representation: a set of individually meaningful topics, and information about which words belong to those topics. We will show that even this simple structure is sufficient to allow the topic model to capture some of the qualitative features of word association that prove problematic for LSA, and to predict quantities that cannot be predicted by LSA, such as the number of meanings or senses of a word.

Comparing topics and spaces

The topic model provides a solution to extracting and using the gist of set of words. In this section, we evaluate the topic model as a psychological account of the content of human semantic memory, comparing its performance with LSA. The topic model and LSA both use the same input – a word-document co-occurrence matrix – but they differ in how this input is analyzed, and in the way that they represent the gist of documents and the meaning of words. By comparing these models, we hope to demonstrate the utility of generative models for exploring questions of semantic representation, and to gain some insight into the strengths and limitations of different kinds of representation.

Our comparison of the topic model and LSA will have two parts. In this section, we analyze the predictions of the two models in depth using a word association task, considering both the quantitative and the qualitative properties of these predictions. In particular, we show that the topic model can explain several phenomena of word association that are problematic for LSA. These phenomena are analogues of the phenomena of similarity judgments that are problematic for spatial models of similarity (Tversky, 1977; Tversky & Gati, 1982; Tversky & Hutchinson, 1986). In the next section we compare the two models across a broad range of tasks, showing that the topic model produces the phenomena that were originally used to support LSA, and describing how the model can be used to predict different aspects of human language processing and memory.

Quantitative predictions for word association

> Are there any more fascinating data in psychology than tables of association?
>
> Deese (1965, p. viii)

Association has been part of the theoretical armory of cognitive psychologists since Thomas Hobbes used the notion to account for the structure of our "Trayne of Thoughts" (Hobbes, 1651/1998; detailed histories of association are provided by Deese, 1965, and Anderson & Bower, 1974). One of the first experimental studies of association was conducted by Galton (1880), who used a word association task to study different kinds of association. Since Galton, several psychologists have tried to classify kinds of association or to otherwise divine its structure (e.g., Deese,

1962; 1965). This theoretical work has been supplemented by the development of extensive word association norms, listing commonly named associates for a variety of words (e.g., Cramer, 1968; Kiss, Armstrong, Milroy, & Piper, 1973; Nelson, McEvoy & Schreiber, 1998). These norms provide a rich body of data, which has only recently begun to be addressed using computational models (Dennis, 2003; Nelson, McEvoy, & Dennis, 2000; Steyvers, Shiffrin, & Nelson, 2004).

While, unlike Deese (1965), we suspect that there may be more fascinating psychological data than tables of associations, word association provides a useful benchmark for evaluating models of human semantic representation. The relationship between word association and semantic representation is analogous to that between similarity judgments and conceptual representation, being an accessible behavior that provides clues and constraints that guide the construction of psychological models. Also, like similarity judgments, association scores are highly predictive of other aspects of human behavior. Word association norms are commonly used in constructing memory experiments, and statistics derived from these norms have been shown to be important in predicting cued recall (Nelson, McKinney, Gee, & Janczura, 1998), recognition (Nelson, McKinney, et al., 1998; Nelson, Zhang, & McKinney, 2001), and false memories (Deese, 1959; McEvoy, Nelson, & Komatsu, 1999; Roediger, Watson, McDermott, & Gallo, 2001). It is not our goal to develop a model of word association, as many factors other than semantic association are involved in this task (e.g., Ervin, 1961; McNeill, 1966), but we believe that issues raised by word association data can provide insight into models of semantic representation.

We used the norms of Nelson et al. (1998) to evaluate the performance of LSA and the topic model in predicting human word association. These norms were collected using a free association task, in which participants were asked to produce the first word that came into their head in response to a cue word. The results are unusually complete, with associates being derived for every word that was produced more than once as an associate for any other word. For each word, the norms provide a set of associates and the frequencies with which they were named, making it possible to compute the probability distribution over associates for each cue. We will denote this distribution $P(w_2|w_1)$ for a cue w_1 and associate w_2, and order associates by this probability: the first associate has highest probability, the second next highest, and so forth.

We obtained predictions from the two models by deriving semantic representations from the TASA corpus (Landauer & Dumais, 1997), which is a collection of excerpts from reading materials commonly encountered between the first year of school and the first year of college. We used a smaller vocabulary than previous applications of LSA to TASA, considering only words that occurred at least 10 times in the corpus and were not included in a standard "stop" list containing function words and other high frequency words with low semantic content. This left us with a vocabulary of 26,243 words, of which 4,235,314 tokens appeared in the 37,651 documents contained in the corpus. We used the singular value decomposition to extract a 700 dimensional representation of the word-document co-occurrence statistics, and examined the performance of the cosine as a predictor of word association using this and a variety of subspaces of lower dimensionality. We also computed the inner product between word vectors as an alternative measure of semantic association, which we will discuss in detail later in the paper. Our choice to use 700 dimensions as an upper limit was guided by two factors, one theoretical and the other practical: previous analyses suggested that the performance of LSA was best with only a few hundred dimensions (Landauer & Dumais, 1997), an observation that was consistent with performance on our task, and 700 dimensions is the limit of standard algorithms for singular value decomposition with a matrix of this size on a workstation with 2GB of RAM.

PRINTING	**PLAY**	TEAM	JUDGE	HYPOTHESIS	STUDY	**CLASS**	ENGINE
PAPER	PLAYS	GAME	TRIAL	EXPERIMENT	**TEST**	MARX	FUEL
PRINT	STAGE	BASKETBALL	**COURT**	SCIENTIFIC	STUDYING	ECONOMIC	ENGINES
PRINTED	AUDIENCE	PLAYERS	CASE	OBSERVATIONS	HOMEWORK	CAPITALISM	STEAM
TYPE	THEATER	PLAYER	JURY	SCIENTISTS	NEED	CAPITALIST	GASOLINE
PROCESS	ACTORS	**PLAY**	ACCUSED	EXPERIMENTS	**CLASS**	SOCIALIST	AIR
INK	DRAMA	PLAYING	GUILTY	SCIENTIST	MATH	SOCIETY	**POWER**
PRESS	SHAKESPEARE	SOCCER	DEFENDANT	EXPERIMENTAL	TRY	SYSTEM	COMBUSTION
IMAGE	ACTOR	PLAYED	JUSTICE	**TEST**	TEACHER	**POWER**	DIESEL
PRINTER	THEATRE	BALL	**EVIDENCE**	METHOD	WRITE	RULING	EXHAUST
PRINTS	PLAYWRIGHT	TEAMS	WITNESSES	HYPOTHESES	PLAN	SOCIALISM	MIXTURE
PRINTERS	PERFORMANCE	BASKET	CRIME	TESTED	ARITHMETIC	HISTORY	GASES
COPY	DRAMATIC	FOOTBALL	LAWYER	**EVIDENCE**	ASSIGNMENT	POLITICAL	CARBURETOR
COPIES	COSTUMES	SCORE	WITNESS	BASED	PLACE	SOCIAL	GAS
FORM	COMEDY	**COURT**	ATTORNEY	OBSERVATION	STUDIED	STRUGGLE	COMPRESSION
OFFSET	TRAGEDY	GAMES	HEARING	SCIENCE	CAREFULLY	REVOLUTION	JET
GRAPHIC	**CHARACTERS**	TRY	INNOCENT	FACTS	DECIDE	WORKING	BURNING
SURFACE	SCENES	COACH	DEFENSE	DATA	IMPORTANT	PRODUCTION	AUTOMOBILE
PRODUCED	OPERA	GYM	CHARGE	RESULTS	NOTEBOOK	CLASSES	STROKE
CHARACTERS	PERFORMED	SHOT	CRIMINAL	EXPLANATION	REVIEW	BOURGEOIS	INTERNAL

Figure 7. A sample of 1700 topics derived from the TASA corpus. Each column contains the 20 highest probability words in a single topic, as indicated by $P(w|z)$. Words in boldface occur in different senses in neighboring topics, illustrating how the model deals with polysemy and homonymy. These topics were discovered in a completely unsupervised fashion, using just word-document co-occurrence frequencies.

We applied the algorithm for finding topics described in Appendix A to the same word-document co-occurrence matrix, extracting representations with up to 1700 topics. Our algorithm is far more memory efficient than the singular value decomposition, as all of the information required throughout the computation can be stored in sparse matrices. Consequently, we ran the algorithm at increasingly high dimensionalities, until prediction performance began to level out. In each case, the set of topics found by the algorithm was highly interpretable, expressing different aspects of the content of the corpus. A selection of topics from the 1700 topic solution are shown in Figure 7.

The topics found by the algorithm pick out some of the key notions addressed by documents in the corpus, including very specific subjects like printing and combustion engines. The topics are extracted purely on the basis of the statistical properties of the words involved – roughly, that these words tend to appear in the same documents – and the algorithm does not require any special initialization or other human guidance. The topics shown in the figure were chosen to be representative of the output of the algorithm, and to illustrate how polysemous and homonymous words are represented in the model: different topics capture different contexts in which words are used, and thus different meanings or senses. For example, the first two topics shown in the figure capture two different meanings of CHARACTERS: the symbols used in printing, and the personas in a play.

To model word association with the topic model, we need to specify a probabilistic quantity that corresponds to the strength of association. The discussion of the problem of prediction above suggests a natural measure of semantic association: $P(w_2|w_1)$, the probability of word w_2 given word w_1. Using the single topic assumption, we have

$$P(w_2|w_1) = \sum_z P(w_2|z)P(z|w_1), \tag{9}$$

which is just Equation 8 with $n = 1$. The details of evaluating this probability are given in Appendix A. This conditional probability automatically compromises between word frequency and semantic relatedness: higher frequency words will tend to have higher probabilities across all topics, and this will be reflected in $P(w_2|z)$, but the distribution over topics obtained by conditioning on w_1, $P(z|w_1)$, will ensure that semantically related topics dominate the sum. If w_1 is highly diagnostic of a particular topic, then that topic will determine the probability distribution over w_2. If w_1 provides no information about the topic, then $P(w_2|w_1)$ will be driven by word frequency.

The overlap between the words used in the norms and the vocabulary derived from TASA was 4,471 words, and all analyses presented in this paper are based on the subset of the norms that uses these words. Our evaluation of the two models in predicting word association was based upon two performance measures: the median rank of the first five associates under the ordering imposed by the cosine or the conditional probability, and the probability of the first associate being included in sets of words derived from this ordering. For LSA, the first of these measures was assessed by computing the cosine for each word w_2 with each cue w_1, ranking the choices of w_2 by $\cos(w_1, w_2)$ such that the highest ranked word had highest cosine, and then finding the ranks of the first five associates for that cue. After applying this procedure to all $4,471$ cues, we computed the median ranks for each of the first five associates. An analogous procedure was performed with the topic model, using $P(w_2|w_1)$ in the place of $\cos(w_1, w_2)$. The second of our measures was the probability that the first associate is included in the set of the m words with the highest ranks under each model, varying m. These two measures are complementary: the first indicates central tendency, while the second gives the distribution of the rank of the first associate.

The topic model outperforms LSA in predicting associations between words. The results of our analyses are shown in Figure 8. We tested LSA solutions with 100, 200, 300, 400, 500, 600 and 700 dimensions. In predicting the first associate, performance levels out at around 500 dimensions, being approximately the same at 600 and 700 dimensions. We will use the 700 dimensional solution for the remainder of our analyses, although our points about the qualitative properties of LSA hold regardless of dimensionality. The median rank of the first associate in the 700 dimensional solution was 31 out of 4470, and the word with highest cosine was the first associate in 11.54% of cases. We tested the topic model with 500, 700, 900, 1100, 1300, 1500, and 1700 topics, finding that performance levels out at around 1500 topics. We will use the 1700 dimensional solution for the remainder of our analyses. The median rank of the first associate in $P(w_2|w_1)$ was 18, and the word with highest probability under the model was the first associate in 16.15% of cases, in both cases an improvement of around 40 percent on LSA.

The performance of both models on the two measures was far better than chance, which would be 2235.5 and 0.02% for the median rank and the proportion correct respectively. The dimensionality reduction performed by the models seems to improve predictions. The conditional probability $P(w_2|w_1)$ computed directly from the frequencies with which words appeared in different documents gave a median rank of 50.5 and predicted the first associate correctly in 10.24% of cases. Latent Semantic Analysis thus improved on the raw co-occurrence probability by between 20 and 40 percent, while the topic model gave an improvement of over 60 percent. In both cases, this improvement results purely from having derived a lower-dimensional representation from the raw frequencies.

Figure 9 shows some examples of the associates produced by people and by the two different models. The figure shows two examples randomly chosen from each of four sets of cues: those for which both models correctly predict the first associate, those for which only the topic model predicts the first associate, those for which only LSA predicts the first associate, and those for which neither model predicts the first associate. These examples help to illustrate how the two models sometimes fail. For example, LSA sometimes latches onto the wrong sense of a word, as with PEN, and tends to give high scores to inappropriate low-frequency words such as WHALE, COMMA, and MILDEW. Both models sometimes pick out correlations between words that do not occur for reasons having to do with the meaning of those words: BUCK and BUMBLE both occur with DESTRUCTION in a single document, which is sufficient for these low frequency words to become associated. In some

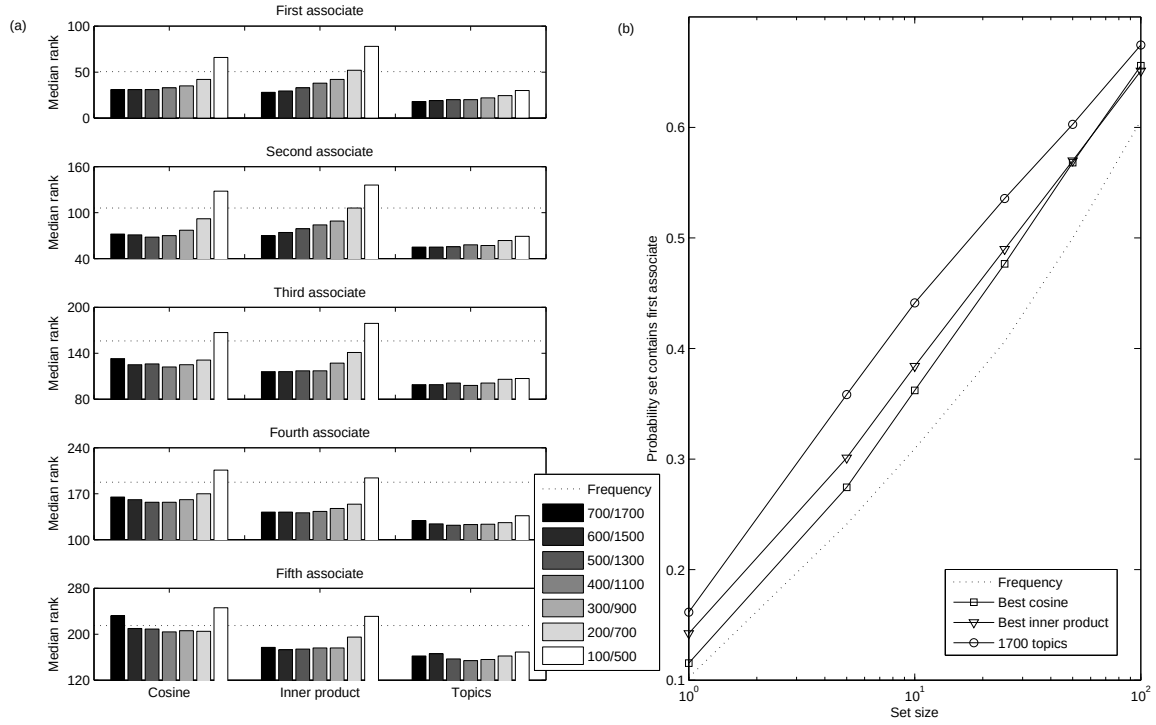

Figure 8. Performance of LSA and the topic model in predicting word association. (a) The median ranks of the first five empirical associates in the ordering predicted by different measures of semantic association at different dimensionalities. Smaller ranks indicate better performance. The dotted line shows baseline performance, corresponding to the use of the raw frequencies with which words occur in the same documents. (b) The probability that a set containing the m highest ranked words under the different measures would contain the first empirical associate, with plot markers corresponding to $m = 1, 5, 10, 25, 50, 100$. The results for the cosine and inner product are the *best* results obtained over all choices of between 100 and 700 dimensions, while the results for the topic model use just the 1700 topic solution. The dotted line is baseline performance derived from co-occurrence frequency.

cases, as with RICE, the most salient properties of an object are not those that are reflected in its use, and the models fail despite producing meaningful, semantically-related predictions.

Qualitative properties of word association

Quantitative measures such as those shown in Figure 8 provide a simple means of summarizing the performance of the two models. However, they mask some of the deeper qualitative differences that result from using different kinds of representations. Tversky (1977; Tversky & Gati, 1982; Tversky & Hutchinson, 1986) argued against defining the similarity between two stimuli in terms of the distance between those stimuli in an internalized spatial representation. Tversky's argument was founded upon violations of the metric axioms – formal principles that hold for all distance measures, which are also known as metrics – in similarity judgments. Specifically, similarity can be asymmetric, since the similarity of x to y can differ from the similarity of y to x, violates the triangle inequality, since x can be similar to y and y to z without x being similar to z, and shows a neighborhood structure inconsistent with the constraints imposed by spatial representations. Tver-

Both		Topics only		LSA only		Neither	
Cue		**Cue**		**Cue**		**Cue**	
CURB	UNCOMMON	PEN	SKILLET	DESTRUCTION	SEPARATE	SPOTS	RICE
Associates		**Associates**		**Associates**		**Associates**	
STREET	COMMON	PENCIL	PAN	DESTROY	DIVIDE	DOG	CHINESE
SIDEWALK	RARE	INK	FRY	WAR	DIVORCE	DIRTY	WEDDING
ROAD	WEIRD	PAPER	EGG	RUIN	PART	DIRT	FOOD
CAR	UNUSUAL	WRITE	COOK	DEATH	SPLIT	STRIPES	WHITE
TIRE	UNIQUE		IRON	KILL	REMOVE	DARK	CHINA
LSA		**LSA**		**LSA**		**LSA**	
STREET	COMMON	HOG	COOKING	DESTROY	DIVIDE	SPOT	PADDY
PEDESTRIAN	FREQUENT	HEN	COOKED	VULNERABLE	INDEPENDENT	GIRAFFE	HARVEST
TRAFFIC	CIRCUMSTANCE	NAP	OVEN	BUMBLE	MIXTURE	GRAY	WHEAT
SIDEWALK	COUPLE	FIX	FRIED	THREAT	ACCOUNT	HIKE	BARLEY
AVENUE	WHALE	MOP	COOK	BOMB	COMMA	MILDEW	BEANS
(1)	(1)	(7)	(6)	(1)	(1)	(2972)	(322)
Topics		**Topics**		**Topics**		**Topics**	
STREET	COMMON	PENCIL	PAN	WAR	FORM	SPOT	VILLAGE
CAR	CASE	FOUNTAIN	KITCHEN	BUCK	SINGLE	FOUND	CORN
CORNER	BIT	INK	COOKING	TAP	DIVISION	GIRAFFE	WHEAT
WALK	EASY	PAPER	STOVE	NUCLEAR	COMMON	BALD	GRAIN
SIDEWALK	KNOWLEDGE	WRITE	POT	DAMAGE	DIVIDE	COVERED	FOOD
(1)	(1)	(1)	(1)	(24)	(5)	(1563)	(68)

Figure 9. Actual and predicted associates for a subset of cues. Two cues were randomly selected from the sets of cues for which (from left to right) both models correctly predicted the first associate, only the topic model made the correct prediction, only LSA made the correct prediction, and neither model made the correct prediction. Each column lists the cue, human associates, predictions of the topic model, and predictions of LSA, presenting the first five words in order. The rank of the first associate is given in parentheses below the predictions of the topic model and LSA.

sky concluded that conceptual stimuli are better represented in terms of sets of features.

Tversky's arguments about the adequacy of spaces and features for capturing the similarity between conceptual stimuli have direct relevance to the investigation of semantic representation. Words are conceptual stimuli, and Latent Semantic Analysis assumes that words can be represented as points in a space. The cosine, the standard measure of association used in LSA, is a monotonic function of the angle between two vectors in a high dimensional space. The angle between two vectors is a metric, satisfying the metric axioms of being zero for identical vectors, being symmetric, and obeying the triangle inequality. Consequently, the cosine exhibits many of the constraints of a metric.

The topic model does not suffer from the same constraints. In fact, the topic model can be thought of as providing a feature-based representation for the meaning of words, with the topics under which a word has high probability being its features. In Appendix B, we show that there is actually a formal correspondence between evaluating $P(w_2|w_1)$ using Equation 9 and computing similarity in one of Tversky's (1977) feature-based models. The association between two words is increased by each topic that assigns high probability to both, and decreased by topics that assign high probability to one but not the other, in the same way that Tverksy claimed common and distinctive features should affect similarity.

The two models we have been considering thus correspond to the two kinds of representation considered by Tversky. Word association also exhibits phenomena that parallel Tversky's analyses of similarity, being inconsistent with the metric axioms. We will discuss three qualitative phenomena of word association – effects of word frequency, violation of the triangle inequality, and the large scale structure of semantic networks – connecting these phenomena to the notions used in Tversky's (1977; Tversky & Gati, 1982; Tversky & Hutchinson, 1986) critique of spatial represen-

tations. We will show that LSA cannot explain these phenomena (at least when the cosine is used as the measure of semantic association), due to the constraints that arise from the use of distances, but that these phenomena emerge naturally when words are represented using topics, just as they can be produced using feature-based representations for similarity.

Asymmetries and word frequency.

The asymmetry of similarity judgments was one of Tversky's (1977) objections to the use of spatial representations for similarity. By definition, any metric d has to be symmetric: $d(x, y) = d(y, x)$. If similarity is a function of distance, similarity should also be symmetric. However, it is possible to find stimuli for which people produce asymmetric similarity judgments. One classic example involves China and North Korea: people typically have the intuition that North Korea is more similar to China than China is to North Korea. Tversky's explanation for this phenomenon appealed to the distribution of features across these objects: our representation of China involves a large number of features, only some of which are shared with North Korea, while our representation of North Korea involves a small number of features, many of which are shared with China.

Word frequency is an important determinant of whether a word will be named as an associate. This can be seen by looking for asymmetric associations: pairs of words w_1, w_2 in which one word is named as an associate of the other much more often than vice versa (i.e. either $P(w_2|w_1) >> P(w_1|w_2)$ or $P(w_1|w_2) >> P(w_2|w_1)$). The effect of word frequency can then be evaluated by examining the extent to which the observed asymmetries can be accounted for by the frequencies of the words involved. We defined two words w_1, w_2 to be associated if one word was named as an associate of the other at least once (i.e. either $P(w_2|w_1)$ or $P(w_1|w_2) > 0$), and assessed asymmetries in association by computing the ratio of cue-associate probabilities for all associated words, $\frac{P(w_2|w_1)}{P(w_1|w_2)}$. Of the 45,063 pairs of associated words in our subset of the norms, 38,744 (85.98%) had ratios indicating a difference in probability of at least an order of magnitude as a function of direction of association. Good examples of asymmetric pairs include KEG-BEER, TEXT-BOOK, TROUSERS-PANTS, MEOW-CAT and COBRA-SNAKE. In each of these cases, the first word elicits the second as an associate with high probability, while the second is unlikely to elicit the first. Of the 38,744 asymmetric associations, 30,743 (79.35%) could be accounted for by the frequencies of the words involved, with the higher frequency word being named as an associate more often.

Latent Semantic Analysis does not predict word frequency effects, including asymmetries in association. The cosine is used as a measure of the semantic association between two words partly because it counteracts the effect of word frequency. The cosine is also inherently symmetric, as can be seen from Equation 1: $\cos(w_1, w_2) = \cos(w_2, w_1)$ for all words w_1, w_2. This symmetry means that the model cannot predict asymmetries in word association without adopting a more complex measure of the association between words (c.f. Krumhansl, 1978; Nosofsky, 1991). In contrast, the topic model can predict the effect of frequency on word association. Word frequency is one of the factors that contributes to $P(w_2|w_1)$. The model can account for the asymmetries in the word association norms. As a conditional probability, $P(w_2|w_1)$ is inherently asymmetric, and the model correctly predicted the direction of $30,905$ (79.77%) of the $38,744$ asymmetric associations, including all of the examples given above. The topic model thus accounted for almost exactly the same proportion of asymmetries as word frequency – the difference was not statistically significant ($\chi^2(1) = 2.08$, $p = 0.149$).

The explanation for asymmetries in word association provided by the topic model is ex-

tremely similar to Tversky's (1977) explanation for asymmetries in similarity judgments. Following Equation 9, $P(w_2|w_1)$ reflects the extent to which the topics in which w_1 appears give high probability to topic w_2. High frequency words tend to appear in more topics than low frequency words. If w_h is a high frequency word and w_l is a low frequency word, w_h is likely to appear in many of the topics in which w_l appears, but w_l will appear in only a few of the topics in which w_h appears. Consequently, $P(w_h|w_l)$ will be large, but $P(w_l|w_h)$ will be small.

Violation of the triangle inequality.

The triangle inequality is another of the metric axioms: for a metric d, $d(x, z) \leq d(x, y) + d(y, z)$. This is referred to as the triangle inequality because if x, y, and z are interpreted as points comprising a triangle, it indicates that no side of that triangle can be longer than the sum of the other two sides. This inequality places strong constraints on distance measures, and strong constraints on the locations of points in a space given a set of distances. If similarity is assumed to be a monotonically decreasing function of distance, then this inequality translates into a constraint on similarity relations: if x is similar to y and y is similar to z, then x must be similar to z. Tversky and Gati (1982) provided several examples where this relationship does not hold. These examples typically involve shifting the features on which similarity is assessed. For instance, taking an example from James (1890), a gas jet is similar to the moon, since both cast light, and the moon is similar to a ball, because of its shape, but a gas jet is not at all similar to a ball.

Word association violates the triangle inequality. A triangle inequality in association would mean that if $P(w_2|w_1)$ is high, and $P(w_3|w_2)$ is high, then $P(w_3|w_1)$ must be high. It is easy to find sets of words that are inconsistent with this constraint. For example ASTEROID is highly associated with BELT, and BELT is highly associated with BUCKLE, but ASTEROID and BUCKLE have little association. Such cases are the rule rather than the exception, as shown in Figure 10 (a). Each of the histograms shown in the figure was produced by selecting all sets of three words w_1, w_2, w_3 such that $P(w_2|w_1)$ and $P(w_3|w_2)$ were greater than some threshold τ, and computing the distribution of $P(w_3|w_1)$. Regardless of the value of τ, there exist a great many triples in which w_1 and w_3 are so weakly associated as not to be named in the norms.

Latent Semantic Analysis cannot explain violations of the triangle inequality. As a monotonic function of the angle between two vectors, the cosine obeys an analogue of the triangle inequality. Given three vectors w_1, w_2, and w_3, the angle between w_1 and w_3 must be less than or equal to the sum of the angle between w_1 and w_2 and the angle between w_2 and w_3. Consequently, $\cos(w_1, w_3)$ must be greater than the cosine of the sum of the $w_1 - w_2$ and $w_2 - w_3$ angles. Using the trigonometric expression for the cosine of the sum of two angles, we obtain the inequality

$$\cos(w_1, w_3) \geq \cos(w_1, w_2)\cos(w_2, w_3) - \sin(w_1, w_2)\sin(w_2, w_3),$$

where $\sin(w_1, w_2)$ can be defined analogously to Equation 1. This inequality restricts the possible relationships between three words: if w_1 and w_2 are highly associated, and w_2 and w_3 are highly associated, then w_1 and w_3 must be highly associated. Figure 10 (b) shows how the triangle inequality manifests in LSA. High values of $\cos(w_1, w_2)$ and $\cos(w_2, w_3)$ induce high values of $\cos(w_1, w_3)$. The implications of the triangle inequality are made explicit in Figure 10 (d): even for the lowest choice of threshold, the minimum value of $\cos(w_1, w_3)$ was above the 97th percentile of cosines between all words in the corpus.

The expression of the triangle inequality in LSA is subtle. It is hard to find triples for which a high value of $\cos(w_1, w_2)$ and $\cos(w_2, w_3)$ induce a high value of $\cos(w_1, w_3)$, although ASTEROID-BELT-BUCKLE is one such example: of the 4470 words in the norms (excluding self associations),

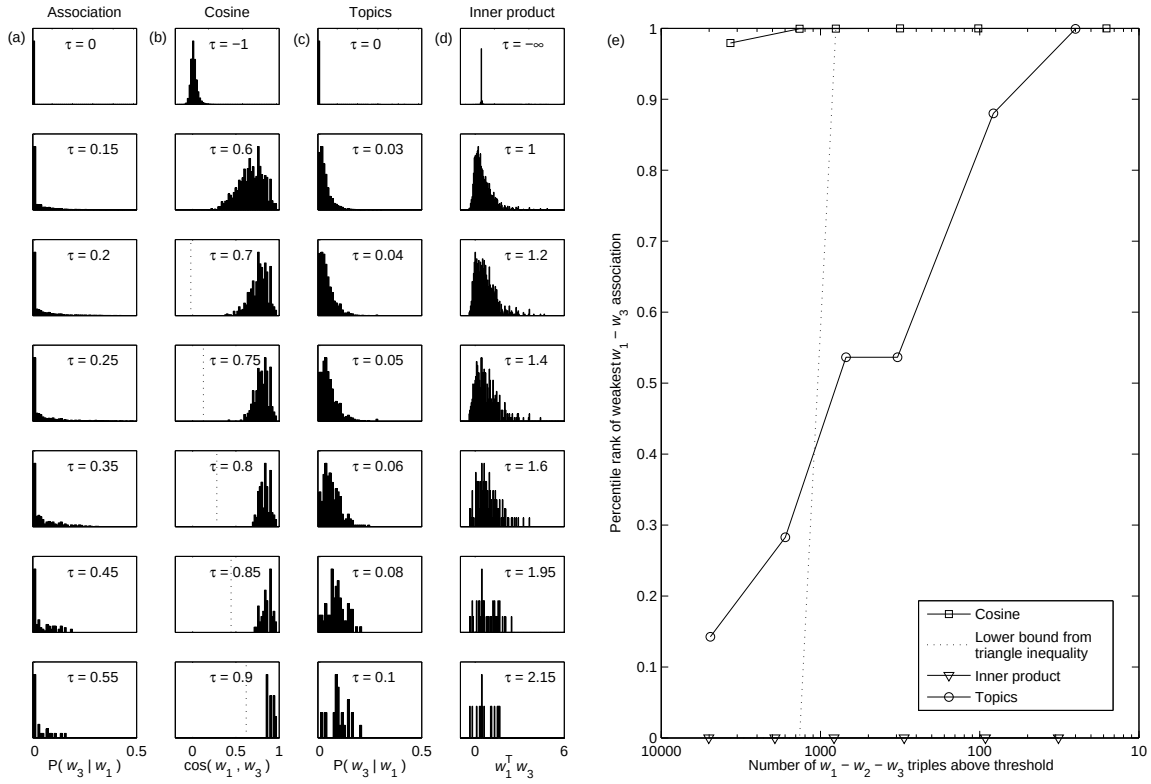

Figure 10. Expression of the triangle inequality in association, Latent Semantic Analysis, and the topic model. (a) Each row gives the distribution of the association probability, $P(w_3|w_1)$, for a triple w_1, w_2, w_3 such that $P(w_2|w_1)$ and $P(w_3|w_2)$ are both greater than τ, with the value of τ increasing down the column. Irrespective of the choice of τ, there remain cases where $P(w_3|w_1) = 0$, suggesting violation of the triangle inequality. (b) Quite different behavior is obtained from LSA, where the triangle inequality enforces a lower bound (shown with the dotted line) on the value of $\cos(w_1, w_3)$ as a result of the values of $\cos(w_2, w_3)$ and $\cos(w_1, w_2)$. (c) The topic model shows only a weak effect of increasing τ, (d) as does the inner product in LSA. In (a)-(d), the value of τ for each plot was chosen to make the number of triples above threshold approximately equal across each row. (e) The significance of the change in distribution can be seen by plotting the percentile rank among all word pairs of the lowest value of $\cos(w_1, w_3)$ and $P(w_3|w_1)$ as a function of the number of triples selected by some value of τ. The plot markers show the percentile rank of the left-most values appearing in the histograms in (b)-(d), for different values of τ. The minimum value of $\cos(w_1, w_3)$ has a high percentile rank even for the lowest value of τ, while $P(w_3|w_1)$ increases more gradually as a function of τ. The minimum inner product remains low for all values of τ.

BELT has the 13th highest cosine with ASTEROID, BUCKLE has the second highest cosine with BELT, and consequently BUCKLE has the 41st highest cosine with ASTEROID, higher than TAIL, IMPACT, or SHOWER. The constraint is typically expressed not by inducing spurious associations between words, but by locating words that might violate the triangle inequality sufficiently far apart that they are unaffected by the limitations it imposes. As shown in Figure 10(b), the theoretical lower bound on $\cos(w_1, w_3)$ only becomes an issue when both $\cos(w_1, w_2)$ and $\cos(w_2, w_3)$ are greater than 0.7.

As illustrated in Figure 7, the topic model naturally recovers the multiple senses of polysemous and homonymous words, placing them in different topics. This makes it possible for violations of the triangle inequality to occur: if w_1 has high probability in topic 1 but not topic 2, w_2 has high probability in both topics 1 and 2, and w_3 has high probability in topic 2 but not topic 1, then $P(w_2|w_1)$ and $P(w_3|w_2)$ can be quite high while $P(w_3|w_1)$ stays low. An empirical demonstration that this is the case for our derived representation is shown in Figure 10 (c): low values of $P(w_3|w_1)$ are observed even when $P(w_2|w_1)$ and $P(w_3|w_2)$ are both high. As shown in Figure 10 (e), the percentile rank of the minimum value of $P(w_3|w_1)$ starts very low, and increases far more slowly than the cosine.

Predicting the structure of semantic networks.

Word association data can be used to construct semantic networks, with nodes representing words and edges representing a non-zero probability of a word being named as an associate. The semantic networks formed in this way can be directed, marking whether a particular word acted as a cue or an associate using the direction of each edge, or undirected, with an edge between words regardless of which acted as the cue. Steyvers and Tenenbaum (2005) analyzed the large scale properties of both directed and undirected semantic networks formed from the word association norms of Nelson et al. (1998), finding that they have some statistical properties that distinguish them from classical random graphs. The properties that we will focus on here are scale-free degree distributions and clustering.

In graph theory, the "degree" of a node is the number of edges associated with that node, equivalent to the number of neighbors. For a directed graph, the degree can differ based on the direction of the edges involved: the in-degree is the number of incoming edges, and the out-degree the number outgoing. By aggregating across many nodes, it is possible to find the degree distribution for a particular graph. Research on networks arising in nature has found that for many such networks the degree k follows a power-law distribution, with $P(k) \sim k^{-\gamma}$ for some constant γ. Such a distribution is often called "scale free", because power-law distributions are invariant with respect to mutiplicative changes of the scale. A power-law distribution can be recognized by plotting $\log P(k)$ against $\log k$: if $P(k) \sim k^{-\gamma}$ then the result should be a straight line with slope $-\gamma$.

Steyvers and Tenenbaum (2005) found that semantic networks constructed from word association data have power-law degree distributions. We reproduced their analyses for our subset of Nelson et al.'s (1998) norms, computing the degree of each word for both directed and undirected graphs constructed from the norms. The degree distributions are shown in Figure 11. In the directed graph, the out-degree (the number of associates for each cue) follows a distribution that is unimodal and exponential-tailed, but the in-degree (the number of cues for which a word is an associate) follows a power-law distribution, indicated by the linearity of $\log P(k)$ as a function of $\log k$. This relationship induces a power-law degree distribution in the undirected graph. We computed three summary statistics for these two power-law distributions: the mean degree, $\bar{k}$, the standard deviation of k, s_k, and the best-fitting power-law exponent, γ. The mean degree serves to describe the overall

Table 1: Structural Statistics and Correlations for Semantic Networks

	Undirected ($\bar{k} = 20.16$)				Directed ($\bar{k} = 11.67$)	
	Association	Cosine	Inner product	Topics	Association	Topics
Statistics						
s_k	18.08	14.51	33.77	21.36	18.72	21.65
γ	2.999	1.972	1.176	2.746	2.028	1.948
$\bar{C}$	0.187	0.267	0.625	0.303	0.187	0.308
$\bar{L}$	3.092	3.653	2.939	3.157	4.298	4.277
Correlations						
k	(0.530)	0.104	0.465	0.487	(0.582)	0.606
C	(-0.462)	0.146	0.417	0.396	(-0.462)	0.391

Note: $\bar{k}$ and s_k are the mean and standard deviation of the degree distribution, γ the power law exponent, $\bar{C}$ the mean clustering coefficient, and $\bar{L}$ the mean length of the shortest path between pairs of words. Correlations in parentheses show the results of using word frequency as a predictor.

density of the graph, while s_k and γ are measures of the rate at which $P(k)$ falls off as k becomes large. If $P(k)$ is strongly positively skewed, as it should be for a power-law distribution, then s_k will be large. The relationship between γ and $P(k)$ is precisely the opposite, with large values of γ indicating a rapid decline in $P(k)$ as a function of k. The values of these summary statistics are given in Table 1.

The degree distribution characterizes the number of neighbors for any given node. A second property of semantic networks, clustering, describes the relationships that hold among those neighbors. Semantic networks tend to contain far more clusters of densely interconnected nodes than would be expected to arise if edges were simply added between nodes at random. A standard measure of clustering (Watts & Strogatz, 1998) is the "clustering coefficient", $\bar{C}$, the mean proportion of the neighbors of a node that are also neighbors of one another. For any node w, this proportion is

$$C_w = \frac{T_w}{\binom{k_w}{2}} = \frac{2T_w}{k_w(k_w - 1)},$$

where T_w is the number of neighbors of w that are neighbors of one another, and k_w is the number of neighbors of w. If a node has no neighbors, C_w is defined to be 1. The clustering coefficient, $\bar{C}$, is computed by averaging C_w over all words w. In a graph formed from word association data, the clustering coefficient indicates the proportion of the associates of a word that are themselves associated. Steyvers and Tenenbaum (2005) found that the clustering coefficient of semantic networks is far greater than that of a random graph. The clustering proportions C_w have been found to be useful in predicting various phenomena in human memory, including cued recall (Nelson, McKinney, et al., 1998), recognition (Nelson et al., 2001), and priming effects (Nelson & Goodmon, 2002), although this quantity is typically referred to as the "connectivity" of a word.

Power-law degree distributions in semantic networks are significant because they indicate that some words have extremely large numbers of neighbors. In particular, the power-law in in-degree indicates that there are a small number of words that appear as associates for a great variety of cues. As Steyvers and Tenenbaum (2005) pointed out, this kind of phenomenon is difficult to

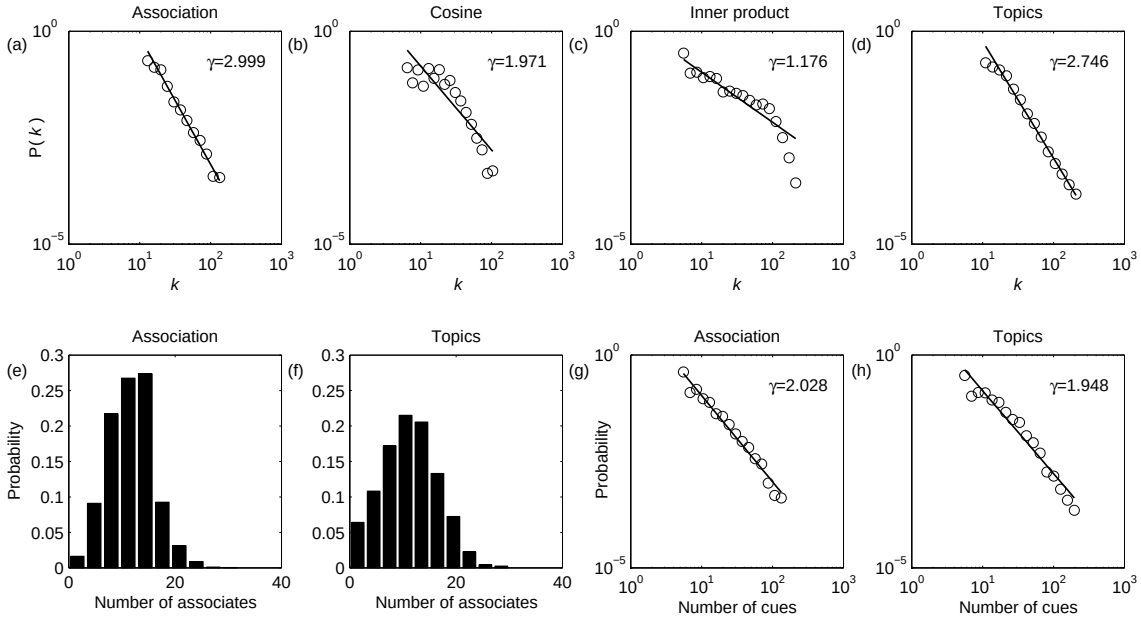

Figure 11. Degree distributions for semantic networks. (a) The power-law degree distribution for the undirected graph, shown as a linear function on log-log coordinates. (b-c) Neither the cosine nor the inner product in LSA produce the appropriate degree distribution. (d) The topic model produces a power-law with the appropriate exponent. (e) In the directed graph, the out-degree is unimodal and exponential-tailed. (f) The topic model produces a similar distribution. (g) The in-degree distribution for the directed graph is power-law. (h) The topic model also provides a close match to this distribution.

reproduce in a spatial representation. This can be demonstrated by attempting to construct the equivalent graph using LSA. Since the cosine is symmetric, the simple approach of connecting each word w_1 to all words w_2 such that $\cos(w_1, w_2) > \tau$ for some threshold τ results in an undirected graph. We used this procedure to construct a graph with the same density as the undirected word association graph, and subjected it to the same analyses. The results of these analyses are presented in Table 1. The degree of individual nodes in the LSA graph is weakly correlated with the degree of nodes in the association graph ($\rho = 0.104$). However, word frequency is a far better predictor of degree ($\rho = 0.530$). Furthermore, the form of the degree distribution is incorrect, as is shown in Figure 11. The degree distribution resulting from using the cosine initially falls off much more slowly than a power-law distribution, resulting in the estimate $\gamma = 1.972$, lower than the observed value of $\gamma = 2.999$, and then falls off more rapidly, resulting in a value of s_k of 14.51, lower than the observed value of 18.08. Similar results are obtained with other choices of dimensionality, and Steyvers and Tenenbaum (2005) found that several more elaborate methods of constructing graphs (both directed and undirected) from LSA were also unable to produce the appropriate degree distribution.

While they exhibit a different degree distribution from semantic networks constructed from association data, graphs constructed by thresholding the cosine seem to exhibit the appropriate amount of clustering. We found C_w for each of the words in our subset of the word association norms, and used these to compute the clustering coefficient $\bar{C}$. We performed the same analysis

on the graph constructed using LSA, and found a similar but slightly higher clustering coefficient. However, LSA differs from the association norms in predicting which words should belong to clusters: the clustering proportions for each word in the LSA graph are only weakly correlated with the corresponding quantities in the word association graph, $\rho = 0.146$. Again, word frequency is a better predictor of clustering proportion, with $\rho = -0.462$.

The neighborhood structure of LSA seems to be inconsistent with the properties of word association. This result is reminiscent of Tversky and Hutchinson's (1986) analysis of the constraints that spatial representations place on the configurations of points in low dimensional spaces. The major concern of Tversky and Hutchinson (1986) was the neighborhood relations that could hold among a set of points, and specifically the number of points to which a point could be the nearest neighbor. In low dimensional spaces, this quantity is heavily restricted: in one dimension, a point can only be the nearest neighbor of two others; in two dimensions, it can be the nearest neighbor of five. This constraint seemed to be at odds with the kinds of structure that can be expressed by conceptual stimuli. One of the examples considered by Tversky and Hutchinson (1986) was hierarchical structure: it seems that apple, orange, and banana should all be extremely similar to the abstract notion of fruit, yet in a low-dimensional spatial representation fruit can only be the nearest neighbor of a small set of points. In word association, power-law degree distributions mean that a few words need to be neighbors of a large number of other words, something that is difficult to produce even in high dimensional spatial representations.

Semantic networks constructed from the predictions of the topic model provide a better match to those derived from word association data. The asymmetry of $P(w_2|w_1)$ makes it possible to construct both directed and undirected semantic networks by thresholding the conditional probability of associates given cues. We constructed directed and undirected graphs by choosing the threshold to match the density, $\bar{k}$, of the semantic network formed from association data. The semantic networks produced by the topic model were extremely consistent with the semantic networks derived from word association, with the statistics are given in Table 1.

As shown in Figure 11 (a), the degree distribution for the undirected graph was power-law with an exponent of $\gamma = 2.746$, and a standard deviation of $s_k = 21.36$, providing a closer match to the true distribution than LSA. Furthermore, the degree of individual nodes in the semantic network formed by thresholding $P(w_2|w_1)$ correlated well with the degree of nodes in the semantic network formed from the word association data, $\rho = 0.487$. The clustering coefficient was close to that of the true graph, $\bar{C} = 0.303$, and the clustering proportions of individual nodes were also well correlated across the two graphs $\rho = 0.396$.

For the directed graph, the topic model produced appropriate distributions for both the out-degree (the number of associates per cue) and the in-degree (the number of cues for which a word is an associate), as shown in Figure 11 (b). The in-degree distribution was power-law, with an exponent of $\gamma = 1.948$ and $s_k = 21.65$, both being close to the true values. The clustering coefficient was similar but slightly higher than the data, $\bar{C} = 0.308$, and the predicted in-degree and clustering proportions of individual nodes correlated well with those for the association graph, $\rho = 0.606$ and $\rho = 0.391$ respectively.

Inner products as an alternative measure of association

In our analyses so far, we have focused on the cosine as a measure of semantic association in LSA, consistent with the vast majority of uses of the model. However, in a few applications, it has been found that the unnormalized inner product gives better predictions (e.g., Rehder et al.,

1998). While it is symmetric, the inner product does not obey a triangle inequality or have easily defined constraints on neighborhood relations. We computed the inner products between all pairs of words from our derived LSA representations, and applied the procedure used to test the cosine and the topic model. We found that the inner product gave better quantitative performance than the cosine, but worse than the topic model, with a median rank for the first associate of 28. A total of 14.23% of the empirical first associates matched the word with the highest inner product. These results are shown in Figure 8. As is to be expected for a measure that does not obey the triangle inequality, there was little effect of the strength of association for (w_1, w_2) pairs and (w_2, w_3) pairs on the strength of association for (w_1, w_3) pairs, as shown in Figure 10 (d) and (e).

As with the other models, we constructed a semantic network by thresholding the inner product, choosing the threshold to match the density of the association graph. The inner product does poorly in reproducing the neighborhood structure of word association, producing a degree distribution that falls off too slowly ($\gamma = 1.176$, $s_k = 33.77$) and an extremely high clustering coefficient ($\bar{C} = 0.625$). However, it does reasonably well in predicting the degree ($\rho = 0.465$) and clustering coefficient ($\rho = 0.417$) of individual nodes. The explanation for this pattern of results is that the inner product is strongly affected by word frequency, and the frequency of words is an important component in predicting associations. However, the inner product gives too much weight to word frequency in forming these predictions, and high frequency words appear as associates for a great many cues. This results in the low exponent and high standard deviation of the degree distribution. The two measures of semantic association used in LSA represent two extremes in their use of word frequency: the cosine is only weakly affected by word frequency, while the inner product is strongly affected. Human semantic memory is sensitive to word frequency, but its sensitivity lies between these extremes.

Summary

The results presented in this section provide analogues in semantic association to the problems that Tversky (1977; Tversky & Gati, 1982; Tversky & Hutchinson, 1986) identified for spatial accounts of similarity. Tversky's argument was not against spatial representations per se, but against the idea that similarity is a monotonic function of a metric, such as distance in psychological space (c.f. Shepard, 1987). Each of the phenomena he noted – asymmetry, violation of the triangle inequality, and neighborhood structure – could be produced from a spatial representation under a sufficiently creative scheme for assessing similarity. Asymmetry provides an excellent example, as several methods for producing asymmetries from spatial representations have already been suggested (Krumhansl, 1978; Nosofsky, 1991). However, his argument shows that the distance between two points in psychological space should not be taken as an absolute measure of the similarity between the objects that correspond to those points. Analogously, our results suggest that the cosine (which is closely related to a metric) should not be taken as an absolute measure of the association between two words.

One way to address some of the problems that we have highlighted in this section may be to use spatial representations in which each word is represented as multiple points, rather than a single point. This is the strategy taken in many connectionist models of semantic representation (e.g., Kawamoto, 1993; Plaut, 1997; Rodd et al., 2004), where different points in space are used to represent different meanings or senses of words. However, typically these representations are not learned from text, but from data consisting of labelled pairs of words and their meanings. Automatically extracting such a representation from text would involve some significant computational

challenges, such as deciding how many senses each word should have, and when those senses are being used.

The fact that the inner product does not exhibit some of the problems we identified with the cosine reinforces the fact that the issue is not with the information extracted by LSA, but with using a measure of semantic association that is related to a metric. The inner product in LSA has an interesting probabilistic interpretation that explains why it should be so strongly affected by word frequency. Under weak assumptions about the properties of a corpus, it can be shown that the inner product between two word vectors is approximately proportional to a smoothed version of the joint probability of those two words (Griffiths & Steyvers, 2003). Word frequency will be a major determinant of this joint probability, and hence has a strong influence on the inner product. This analysis suggests that while the inner product provides a means of measuring semantic association that is nominally defined in terms of an underlying semantic space, much of its success may actually be a consequence of approximating a probability.

The topic model provides an alternative to LSA which automatically solves the problem of understanding the different senses in which a word might be used, and gives a natural probabilistic measure of association that is not subject to the constraints of a metric. It gives more accurate quantitative predictions of word association data than using either the cosine or the inner product in the representation extracted by LSA. It also produces predictions that are consistent with the qualitative properties of semantic association that are problematic for spatial representations. In the remainder of the paper, we consider some further applications of this model, including other comparisons to LSA, and how it can be extended to accommodate more complex semantic and syntactic structures.

Further applications

Our analysis of word association provided an in-depth exploration of the differences between LSA and the topic model. However, these models are intended to provide an account of a broad range of empirical data, collected through a variety of tasks that tap the representations used in processing language. In this section, we present a series of examples of applications of these models to other tasks. These examples show that the topic model reproduces many of the phenomena that were originally used to support LSA, provide a broader basis for comparison between the two models, and illustrate how the representation extracted by the topic model can be used in other settings.

Synonym tests

One of the original applications of LSA was to the TOEFL synonym test, used to assess fluency in English for non-native speakers (Landauer & Dumais, 1997). To allow direct comparison between the predictions of LSA and the topic model, we replicated these results and evaluated the performance of the topic model on the same task. The test contained 90 questions, consisting of a probe word and four answers. Our analyses only included questions for which all five words (probe and answers) were in our 26,243 word vocabulary, resulting in a set of 44 questions. We used the solutions obtained from the TASA corpus, as described in the previous section. For LSA, we computed the cosine and inner product betwen probe and answers for LSA solutions with between 100 and 700 dimensions. For the topic model, we computed $P(w_{probe}|w_{answer})$ and $P(w_{answer}|w_{probe})$ for between 500 and 1700 topics, where w_{probe} and w_{answer} are the probe and answer words respectively, and Equation 8 was used to calculate the conditional probabilities.

Our first step in evaluating the models was to examine how often the answer that each model identified as being most similar to the probe was the correct answer. Landauer and Dumais (1997) reported that LSA (trained on the TASA corpus, but with a larger vocabulary than we used here) produced 64.4% correct answers, close to the average of 64.5% produced by college applicants from non-English-speaking countries. Our results were similar: the best performance using the cosine was with a solution using 500 dimensions, resulting in 63.6% correct responses. There were no systematic effects of number of dimensions, and only a small amount of variation. The inner product likewise produced best performance with 500 dimensions, getting 61.5% correct.

The topic model performed similarly to LSA on the TOEFL test: using $P(w_{probe}|w_{answer})$ to select answers, the best performance was obtained with 500 topics, being 70.5% correct. Again, there was no systematic effect of number of topics. Selecting answers using $P(w_{answer}|w_{probe})$ produced results similar to the cosine for LSA, with the best performance being 63.6% correct, obtained with 500 topics. The difference between these two ways of evaluating the conditional probability lies in whether the frequencies of the possible answers are taken into account. Computing $P(w_{probe}|w_{answer})$ controls for the frequency with which the words w_{answer} generally occur, and is perhaps more desirable in the context of a vocabulary test.

As a final test of the two models, we computed the correlation between their predictions and the actual frequencies with which people selected the different responses. For the LSA solution with 500 dimensions, the mean correlation between the cosine and response frequencies (obtained by averaging across items) was $r = 0.30$, with $r = 0.25$ for the inner product. For the topic model with 500 topics, the corresponding correlations were $r = 0.46$ and 0.34 for $\log P(w_{probe}|w_{answer})$ and $\log P(w_{answer}|w_{probe})$ respectively. Thus, these models produced predictions that were not just correct, but captured some of the variation in human judgments on this task.

Semantic priming of different word meanings

Till, Mross, and Kintsch (1988) examined the time-course of the processing of word meanings using a priming study in which participants read sentences containing ambiguous words and then performed a lexical decision task. The sentences were constructed to provide contextual information about the meaning of the ambiguous word. For example, two of the sentences used in the study were

> 1A. The townspeople were amazed to find that all of the buildings had collapsed except the mint. Obviously, it has been built to withstand natural disasters.
>
> 1B. Thinking of the amount of garlic in his dinner, the guest asked for a mint. He soon felt more comfortable socializing with the others.

which are intended to pick out the different meanings of MINT. The target words used in the lexical decision task corresponded either to the different meanings of the ambiguous word (in this case being MONEY and CANDY), or were inferentially related to the content of the sentence (in this case being EARTHQUAKE and BREATH). The delay between the presentation of the sentence and the decision task was varied, making it possible to examine how the timecourse of processing affected the facilitation of lexical decisions (i.e. priming) for different kinds of targets.

The basic result reported by Till et al. (1988) was that both of the meanings of the ambiguous word and neither of the inference targets were primed when there was a short delay between sentence presentation and lexical decision, and that there was a subsequent shift to favor the appropriate

Table 2: Predictions of models for semantic priming task of Till et al. (1988)

	Meaning A e.g. MONEY	*Inference A* e.g. EARTHQUAKE	*Meaning B* e.g. CANDY	*Inference B* e.g. BREATH
Cosine				
early	0.099	0.038	0.135	0.028
late A	0.060	0.103	0.046	0.017
late B	0.050	0.024	0.067	0.046
Inner product				
early	0.208	0.024	0.342	0.017
late A	0.081	0.039	0.060	0.012
late B	0.060	0.009	0.066	0.024
Topics ($\log_{10}$ *probability)*				
early	-3.22	-4.31	-3.16	-4.42
late A	-4.03	-4.13	-4.58	-4.77
late B	-4.52	-4.73	-4.21	-4.24

meaning and infentially related target when the delay was increased. Landauer and Dumais (1997) suggested that this effect could be explained by LSA, using the cosine between the ambiguous word and the targets to model priming at short delays, and the cosine between the entire sentence and the targets to model priming at long delays. They showed that effects similar to those reported by Till et al. (1988) emerged from this analysis.

We reproduced the analysis of Landauer and Dumais (1997) using the representations we extracted from the TASA corpus. Of the 28 pairs of sentences used by Till et al. (1988), there were 20 for which the ambiguous primes and all four target words appeared in our vocabulary. To simulate priming early in processing, we computed the cosine and inner product between the primes and the target words using the representation extracted by LSA. To simulate priming in the later stages of processing, we computed the cosine and inner product between the average vectors for each of the full sentences (including only those words that appeared in our vocabulary) and the target words. The values produced by these analyses were then averaged over all 20 pairs. The results for the 700 dimensional solution are shown in Table 2 (similar results were obtained with different numbers of dimensions).

The results of this analysis illustrate the trends identified by Landauer and Dumais (1997). Both the cosine and the inner product give reasonably high scores to the two meanings when just the prime is used (relative to the distributions shown in Figure 10), and shift to give higher scores to the meaning and inferentially related target appropriate to the sentence when the entire sentence is used. To confirm that the topic model makes similar predictions in the context of semantic priming, we used the same procedure with the topic-based representation, computing the conditional probabilities of the different targets based just on the prime, and based on the entire sentences, then averaging the log probabilities over all pairs of sentences. The results for the 1700 topic solution are shown in Table 2 (similar results were obtained with different numbers of topics). The topic model produces the same trends: it initially gives high probability to both meanings, and then switches to give high probabilities to the sentence-appropriate targets.

Sensitivity of reading time to frequency of meanings

Examining the time that people take to read words and sentences has been one of the most widely used methods for evaluating the contributions of semantic representation to linguistic processing. In particular, several studies have used reading time to explore the representation of ambiguous words (e.g., Duffy, Morris, & Rayner, 1988; Rayner & Duffy, 1986; Rayner & Frazier, 1989). Developing a complete account of how the kind of contextual information we have been discussing influences reading time is beyond the scope of this paper. However, we used the topic model to predict the results of one such study, to provide an illustration of how it can be applied to a task of this kind.

Sereno, Pacht, and Rayner (1992) conducted a study in which the eye movements of participants were monitored while they read sentences containing ambiguous words. These ambiguous words were selected to have one highly dominant meaning, but the sentences established a context that supported the subordinate mearning. For example, one sentence read

> The dinner party was proceeding smoothly when, just as Mary was serving the port, one of the guests had a heart attack.

where the context supported the subordinate meaning of PORT. The aim of the study was to establish whether reading time for ambiguous words was better explained by the overall frequency with which a word occurs in all its meanings or senses, or the frequency of a particular meaning. To test this, participants read sentences containing either the ambiguous word, a word with frequency matched to the subordinate sense (the low-frequency control), or a word with frequency matched to the dominant sense (the high-frequency control). For example, the control words for PORT were VEAL and SOUP respectively. The results are summarized in Table 3: ambiguous words using their subordinate meaning were read more slowly than words with a frequency corresponding to the dominant meaning, although not quite as slowly as words that match the frequency of the subordinate meaning. A subsequent study by Sereno, O'Donnell, and Rayner (2006, Experiment 3) produced the same pattern of results.

Reading time studies present a number of challenges for computational models. The study of Sereno et al. (1992) is particularly conducive to modeling, as all three target words are substituted into the same sentence frame, meaning that the results are not affected by sentences differing in the number of words in the vocabulary of the models or other factors that introduce additional variance. However, in order to model these data we still need to make an assumption about the factors influencing reading time. The abstract computational-level analyses provided by generative models do not make assertions about the algorithmic processes underlying human cognition, and can consequently be difficult to translate into predictions about the amount of time it should take to perform a task. In the topic model, there are a variety of factors that could produce an increase in the time taken to read a particular word. Some possible candidates include uncertainty about the topic of the sentence, as reflected in the entropy of the distribution over topics, a sudden change in perceived meaning, producing a difference in the distribution over topics before and after seeing the word, or simply encountering an unexpected word, resulting in greater effort for retrieving the relevant information from memory. We chose to use only the last of these measures, being the simplest and the most directly related to our construal of the computational problem underlying linguistic processing, but suspect that a good model of reading time would need to incorporate some combination of all of these factors.

Table 3: Predictions of models for reading time task of Sereno et al. (1992)

	Ambiguous word	Low-frequency control	High-frequency control
Human gaze duration (ms)	281	287	257
Cosine	0.021	0.048	0.043
Inner product	0.011	0.010	0.025
Topics ($\log_{10}$ probability)	-4.96	-5.26	-4.68

Letting w_{target} be the target word and $\mathbf{w}_{sentence}$ be the sequence of words in the sentence before the occurrence of the target, we want to compute $P(w_{target}|\mathbf{w}_{sentence})$. Applying Equation 8, we have

$$P(w_{target}|\mathbf{w}_{sentence}) = \sum_z P(w_{target}|z)P(z|\mathbf{w}_{sentence}) \quad (10)$$

where $P(z|\mathbf{w}_{sentence})$ is the distribution over topics encoding the gist of $\mathbf{w}_{sentence}$. We used the 1700 topic solution to compute this quantity for the 21 of the 24 sentences used by Sereno et al. (1992) for which all three target words appeared in our vocabulary, and averaged the resulting log probabilities over all sentences. The results are shown in Table 3. The topic model predicts the results found by Sereno et al. (1992): the ambiguous words are assigned lower probabilities than the high-frequency controls, although not quite as low as the low-frequency controls. The model predicts this effect because the distribution over topics $P(z|\mathbf{w}_{sentence})$ favors those topics that incorporate the subordinate sense. As a consequence, the probability of the target word is reduced, since $P(w_{target}|z)$ is lower for those topics. However, if there is any uncertainty, providing some residual probability to topics in which the target word occurs in its dominant sense, the probability of the ambiguous word will be slightly higher than the raw frequency of the subordinate sense suggests.

For comparison, we computed the cosine and inner product for the three values of w_{target} and the average vectors for $\mathbf{w}_{sentence}$ in the 700 dimensional LSA solution. The results are shown in Table 3. The cosine does not predict this effect, with the highest mean cosines being obtained by the control words, with little effect of frequency. This is due to the fact that the cosine is relatively insensitive to word frequency, as discussed above. The inner product, which is sensitive to word frequency, produces predictions that are consistent with the results of Sereno et al. (1992).

Semantic intrusions in free recall

Word association involves making inferences about the semantic relationships among a pair of words. The topic model can also be used to make predictions about the relationships between multiple words, as might be needed in episodic memory tasks. Since Bartlett (1932), many memory researchers have proposed that episodic memory might not only be based on specific memory of the experiences episodes but also on reconstructive processes that extract the overall theme or gist of a collection of experiences.

One procedure for studying gist-based memory is the Deese-Roediger-McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995). In this paradigm, participants are instructed to remember short lists of words that are all associatively related to a single word (the critical lure) that is not presented on the list. For example, one DRM list consists of the words BED, REST, AWAKE, TIRED, DREAM, WAKE, SNOOZE, BLANKET, DOZE, SLUMBER, SNORE, NAP,

PEACE, YAWN, and DROWSY. At test, 61% of subjects falsely recall the critical lure SLEEP, which is associatively related to all the presented words.

The topic model may be able to play a part in a theoretical account for these semantic intrusions in episodic memory. Previous theoretical accounts of semantic intrusions have been based on "dual route" models of memory. These models distinguish between different routes to retrieve information from memory, a verbatim memory route based on the physical occurrence of an input and the gist memory route that is based on semantic content (e.g., Brainerd et al., 1999; Brainerd et al., 2002; Mandler, 1980). The representation of the gist or the processes involved in computing the gist itself have not been specified within the dual route framework. Computational modeling in this domain has been mostly concerned with the estimation of the relative strength different memory routes within the framework of multinomial processing tree models (Batchelder & Riefer, 1999).

The topic model can provide a more precise theoretical account of gist-based memory by detailing both the representation of the gist and the inference processes based on the gist. We can model the retrieval probability of a single word at test based on a set of studied words by computing $P(w_{recall}|\mathbf{w}_{study})$. With the topic model, we can use Equation 8 to obtain

$$P(w_{recall}|\mathbf{w}_{study}) = \sum_z P(w_{recall}|z)P(z|\mathbf{w}_{study}). \tag{11}$$

The gist of the study list is represented by $P(z|\mathbf{w}_{study})$ which describes the distribution over topics for a given study list. In the DRM paradigm, each list of words will lead to a different distribution over topics. Lists of relatively unrelated words will lead to flat distributions over topics where no topic is particularly likely, whereas more semantically focused lists will lead to distributions where only a few topics dominate. The term $P(w_{recall}|z)$ captures the retrieval probability of words given each of the inferred topics.

We obtained predictions from this model for the 55 DRM lists reported by Roediger et al. (2001), using the 1700 topic solution derived from the TASA corpus. Three DRM lists were excluded because the critical items were absent from the vocabulary of the model. Of the remaining 52 DRM lists, a median of 14 out of 15 original study words were in our vocabulary. For each DRM list, we computed the retrieval probability over the whole 26,243 word vocabulary which included the studied words as well as extra-list words. For example, Figure 12 shows the predicted gist-based retrieval probabilities for the SLEEP list. The retrieval probabilities are separated into two lists: the words on the study list and the 8 most likely extra-list words. The results shows that the word SLEEP is the most likely word to be retrieved which qualitatively fits with the observed high false recall rate of this word.

To assess the performance of the topic model, we correlated the retrieval probability of the critical DRM words as predicted by the topic model with the observed intrusion rates reported by Roediger et al. (2001). The rank-order correlation was 0.437 with a 95% confidence interval (estimated by 1000 sample bootstrap) of $(0.217, 0.621)$. We compared this performance with the predictions of the 700-dimensional LSA solution. Using LSA, the gist of the study list was represented by the average of all word vectors from the study list. We then computed the cosine of the critical DRM word with the average word vector for the DRM list and correlated this cosine with the observed intrusion rate. The correlation was 0.295, with a 95% confidence interval (estimated by 1000 sample bootstrap) of $(0.041, 0.497)$. The improvement in predicting semantic intrusions produced by the topic model over LSA is thus not statistically significant, but suggests that the two models might be discriminated through further experiments.

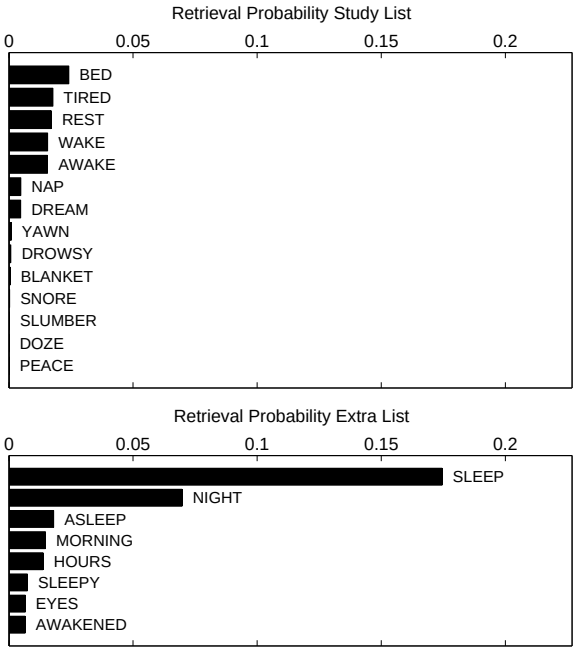

Figure 12. Retrieval probabilities, $P(w_{recall}|\mathbf{w}_{study})$, for a study list containing words semantically associated with SLEEP. The upper panel shows the probabilities of each of the words on the study list. The lower panel shows the probabilities of the most likely extra-list words. SLEEP has a high retrieval probability, and would thus be likely to be falsely recalled.

One interesting observation from Figure 12 is that words that do not appear on the study list, such as SLEEP, can be given higher probabilities than the words that actually do appear on the list. Since participants in free recall studies generally do well in retrieving the items that appear on the study list, this illustrates that the kind of gist-based memory that the topic model embodies is not sufficient to account for behavior on this task. The gist-based retrieval process would have to be complemented with a verbatim retrieval process in order to account for the relatively high retrieval probability for words on the study list, as assumed in the dual-route models mentioned above (Brainerd et al., 1999; Brainerd et al., 2002; Mandler, 1980). These issues could be addressed by extending the topic model to take into account the possible interaction between the gist and verbatim routes.

Meanings, senses, and topics

The topic model assumes a simple structured representation for words and documents, in which words are allocated to individually interpretable topics. This representation differs from that assumed by LSA, in which the dimensions are not individually interpretable and the similarity between words is invariant with respect to rotation of the axes. The topic-based representation also provides the opportunity to explore questions about language that cannot be posed using less structured representations. As we have seen already, different topics can capture different meanings or senses of a word. As a final test of the topic model, we examined how well the number of topics in which a word participates predicts its the number of meanings or senses of that word, and how this quantity can be used in modeling recognition memory.

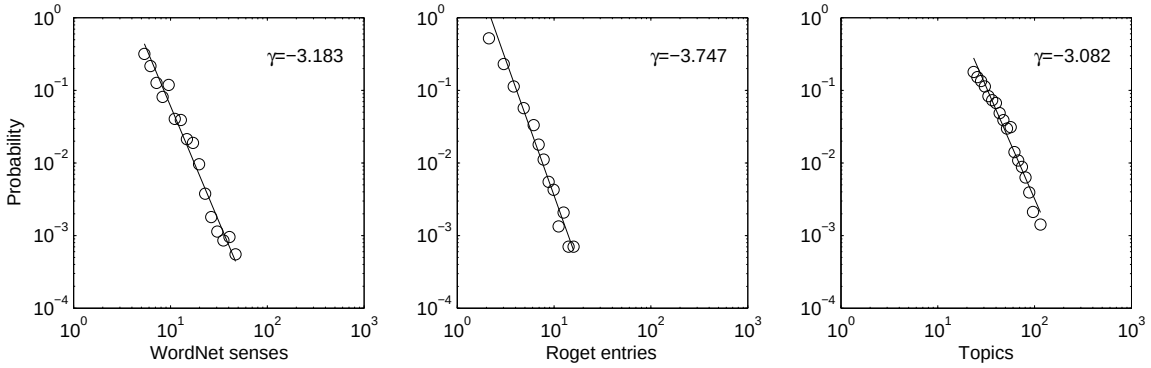

Figure 13. The distribution of the number of contexts in which a word can appear has a characteristic form, whether computed from the number of senses in WordNet, the number of entries in Roget's thesaurus, or the number of topics in which a word appears.

The number of meanings or senses that a word possesses has a characteristic distribution, as was first noted by Zipf (1965). Zipf examined the number of entries that appeared in dictionary definitions for words, and found that this quantity followed a power-law distribution. Steyvers and Tenenbaum (2005) conducted similar analyses using Roget's (1911) thesaurus and WordNet (Miller & Fellbaum, 1998). They also found that the number of entries followed a power-law distribution, with an exponent of $\gamma \approx 3$. Plots of these distributions in log-log coordinates are shown in Figure 13.

The number of topics in which a word appears in the topic model corresponds well with the number of meanings or senses of words as assessed using Roget's thesaurus and WordNet, both in distribution and in the values for individual words. The distribution of the mean number of topics to which a word was assigned in the 1700 topic solution is shown in Figure 13.[4] The tail of this distribution matches the tail of the distributions obtained from Roget's thesaurus and WordNet, with all three distributions being power-law with a similar parameter. Furthermore, the number of topics in which a word appears is closely correlated with these other measures: the rank-order correlation between number of topics and number of entries in Roget's thesaurus is $\rho = 0.328$, with a 95% confidence interval (estimated by 1000 sample bootstrap) of $(0.300, 0.358)$, and the correlation between number of topics and WordNet senses give $\rho = 0.508$, with a 95% confidence interval of $(0.486, 0.531)$. For comparison, the most obvious predictor of the number of meanings or senses of a word – word frequency – gives correlations that fall below these confidence intervals: word frequency predicts Roget entries with a rank-order correlation of $\rho = 0.243$, and WordNet senses with $\rho = 0.431$. More details of the factors affecting the distribution of the number of topics per word are given in Griffiths and Steyvers (2002).

Capturing context variability

The number of topics in which a word appears also provides a novel means of measuring an important property of words: context variability. Recent research in recognition memory has

[4] As the number of topics to which a word is assigned will be affected by the number of topics in the solution, these values cannot be taken as representing the number of meanings or senses of a word directly. As mentioned previously, the correspondence will be many-to-one.

Table 4: Correlations of recognition memory sensitivities (d') with word frequency (WF), document frequency (DF) and topic variability (TV).

Variable	d'	$\log WF$	$\log DF$
$\log WF$	-0.50^*	-	-
$\log DF$	-0.58^*	0.97^*	-
$\log TV$	-0.67^*	0.69^*	0.82^*
$\log TV \mid \log WF^{**}$	-0.53^*	-	-
$\log TV \mid \log DF^{**}$	-0.43^*	-	-

Note: ($*$) Correlations are significant at $p < .0001$. ($**$) Partial correlations where the effect of the second variable is partialled out of the effect of the first variable.

suggested that the number of contexts in which words appear might explain why some words are more likely than others to be confused for items appearing on the study list in recognition memory experiments (Dennis & Humphreys, 2001; McDonald & Shillcock, 2001; Steyvers & Malmberg, 2003). The explanation for this effect is that when a word is encountered in a larger number of contexts, the study list context becomes less discriminable from these previous exposures (Dennis & Humphreys, 2001). Steyvers and Malmberg (2003) operationally defined context variability as the number of documents in which a word appears in a large database of text, a measure we will refer to as *document frequency*. Steyvers and Malmberg found that this measure has an effect on recognition memory independent of word frequency. The document frequency measure is a rough proxy for context variability because it does not take the other words occurring in documents into account. The underlying assumption is that documents are equally different from each other. Consequently, if there are many documents that cover very similar sets of topics, then context variability will be overestimated.

The topic model provides an alternative way to assess context variability. Words that are used in different contextual uses tend to be associated with different topics. Therefore, we can assess context variability by the number of different topics a word is associated with, a measure we will refer to as *topic variability*. Unlike document frequency, this measure does take into account the similarity between different documents in evaluating context variability.

To understand how topic variability compares with word frequency and contextual variability, we performed analyses on the data from the experiment by Steyvers and Malmberg (2003). There were 287 distinct words in the experiment each being used as either a target or a distractor. For each word, we computed the sensitivity (d') measuring the degree to which subjects could distinguish that word as a target or distractor in the recognition memory experiment. Table 4 shows the correlations between d' and the three measures: topic variability (TV), word frequency (WF) and document frequency (DF). All three word measures were logarithmically scaled.

The results show that word frequency, context variability, and topic variability all correlate with recognition memory performance as expected – high word frequency, high document frequency, and high topic variability are all associated with poor recognition memory performance. Topic variability correlates more strongly with performance than the other measures ($p < 0.05$) and is also less correlated with the other measures. This suggests that topic variability is a good predictive measure for recognition memory confusability and is at least as good a predictor as word frequency or document frequency, and potentially a more direct measure of context variability.

Summary

The results presented in this section illustrate that the topic model can be used to predict behavior on a variety of tasks relating to linguistic processing and semantic memory. The model reproduces many of the phenomena that have been used to support Latent Semantic Analysis, and consistently provides better performance than using the cosine or the inner product between word vectors to measure semantic association. The form of the representation extracted by the topic model also makes it possible to define novel measures of properties of words such as the number of topics in which they appear, which seems to be a good guide to the number of senses or meanings of a word, as well as an effective predictor of recognition memory performance.

Extending the generative model

Formulating the problem of extracting and using gist in terms of generative models allowed us to explore a novel form of semantic representation, through the topic model. This formulation of the problem also has other advantages. Generative models provide a very flexible framework for specifying structured probability distributions, and it is easy to extend the topic model to incorporate richer latent structure by adding further steps to the generative process. We will discuss five extensions to the model: determining the number of topics, learning topics from other kinds of data, incorporating collocations, inferring topic hierarchies, and including rudimentary syntax.

Learning the number of topics

In the preceding discussion, we assumed that the number of topics, T, in the model was fixed. This assumption seems inconsistent with the demands of human language processing, where more topics are introduced with every conversation. Fortunately, this assumption is not necessary. Using methods from non-parametric Bayesian statistics (Muller & Quintana, 2004; Neal, 2000), we can assume that our data are generated by a model with an unbounded number of dimensions, of which only a finite subset have been observed. The basic idea behind these non-parametric approaches is to define a prior probability distribution on the assignments of words to topics, $\mathbf{z}$, that does not assume an upper bound on the number of topics. Inferring the topic assignments for the words that appears in a corpus simultaneously determines the number of topics, as well as their content. Blei, Griffiths, Jordan, and Tenenbaum (2004) and Teh, Jordan, Beal, and Blei (2004) have applied this strategy to learn the dimensionality of topic models. These methods are closely related to the rational model of categorization proposed by Anderson (1990), which represents categories in terms of a set of clusters, with new clusters being added automatically as more data becomes available (see Neal, 2000).

Learning topics from other data

Our formulation of the basic topic model also assumes that words are divided into documents, or otherwise broken up into units that share the same gist. A similar assumption is made by LSA, but this is not true of all methods for automatically extracting semantic representations from text (e.g., Dennis, 2004; Lund & Burgess, 1996; Jones & Mewhort, 2006). This assumption is not appropriate for all settings in which we make linguistic inferences: while we might differentiate the documents we read, many forms of linguistic interaction, such as meetings or conversations, lack clear markers that break them up into sets of words with a common gist. One approach to this problem is to define a generative model in which the document boundaries are also latent variables, a strategy pursued

by Purver, Koerding, Griffiths, and Tenenbaum (2006). Alternatively, meetings or conversations might be better modeled by associating the gist of a set of words with the person who utters those words, rather than words in temporal proximity. Rosen-Zvi, Griffiths, Steyvers, and Smyth, 2004) and Steyvers, Smyth, Rosen-Zvi, and Griffiths (2004) have extensively investigated models of this form.

Inferring topic hierarchies

We can also use the generative model framework as the basis for defining models that use richer semantic representations. The topic model assumes that topics are chosen independently when generating a document. However, people know that topics bear certain relations to one another, and that words have relationships that go beyond topic membership. For example, some topics are more general than others, subsuming some of the content of those other topics. The topic of sport is more general than the topic of tennis, and the word SPORT has a wider set of associates than TENNIS. These issues can be addressed by developing models in which the latent structure concerns not just the set of topics that participate in a document, but the relationships among those topics. Generative models that use topic hierarchies provide one example of this, making it possible to capture the fact that certain topics are more general than others. Blei, Griffiths, Jordan and Tenenbaum (2004) provided an algorithm that simultaneously learns the structure of a topic hierarchy, and the topics that are contained within that hierarchy. This algorithm can be used to extract topic hierarchies from large document collections. Figure 14 shows the results of applying this algorithm to the abstracts of all papers published in Psychological Review since 1967. The algorithm recognizes that the journal publishes work in cognitive psychology,[5] social psychology, vision research, and biopsychology, splitting these subjects into separate topics at the second level of the hierarchy, and finds meaningful subdivisions of those subjects at the third level. Similar algorithms can be used to explore other representations that assume dependencies among topics (Blei & Lafferty, 2006).

Collocations and associations based on word order

In the basic topic model, the probability of a sequence of words is not affected by the order in which they appear. As a consequence, the representation extracted by the model can only capture coarse-grained contextual information, such as the fact that words tend to appear in the same sort of conversations or documents. This is reflected in the fact that the input to the topic model, as with LSA, is a word-document co-occurrence matrix: the order in which the words appear in the documents does not matter. However, it is clear that word order is important to many aspects of linguistic processing, including the simple word association task that we discussed extensively earlier in the paper (Ervin, 1961; Hutchinson, 2003; McNeill, 1966).

A first step towards relaxing the insensitivity to word order displayed by the topic model is to extend the model to incorporate collocations – words that tend to follow one another with high frequency. For example, the basic topic model would treat the phrase UNITED KINGDOM occurring in a document as one instance of UNITED and one instance of KINGDOM. However, these two words carry more semantic information when treated as a single chunk than they do alone. By extending the model to incorporate a sensitivity to collocations, we also have the opportunity to examine how incorporating this additional source of predictive information affects predictions about the associations that exist between words.

[5] Or, more precisely, psychology based upon an information-processing approach to studying the mind.

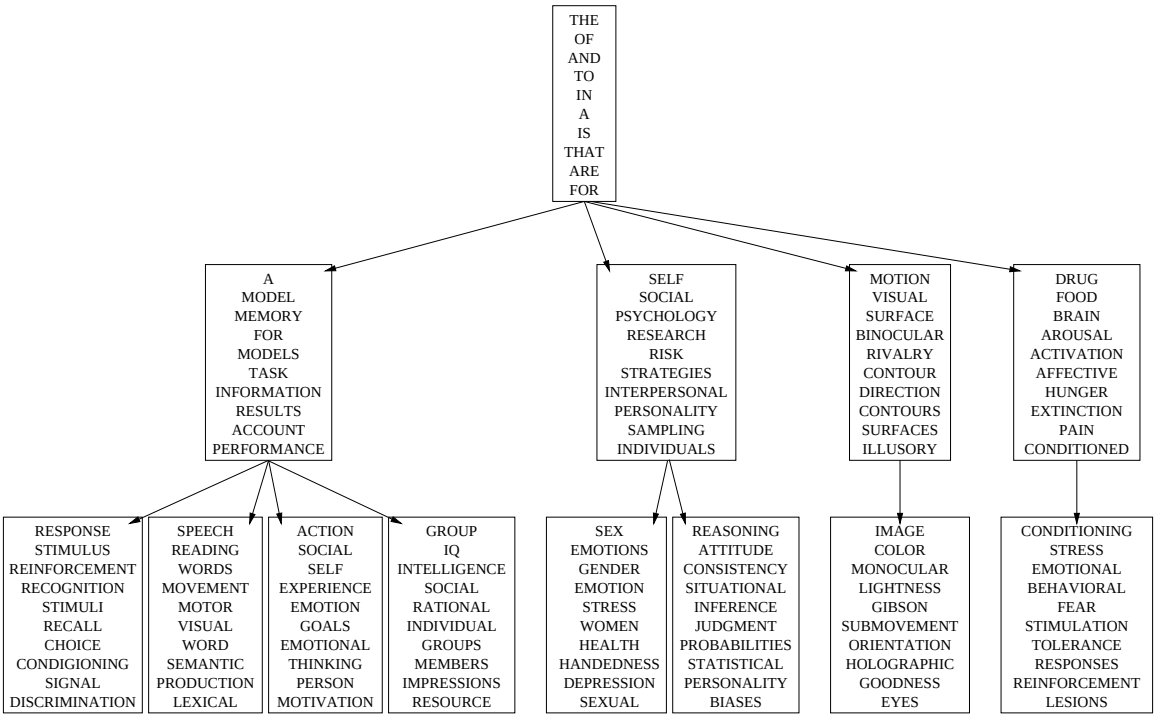

Figure 14. A topic hierarchy, learned from the abstracts of articles appearing in *Psychological Review* since 1967. Each document is generated by choosing a path from the root (the top node) to a leaf (the bottom nodes). Consequently, words in the root topic appear in all documents, the second level topics pick out broad trends across documents, and the topics at the leaves pick out specific topics within those trends. The model differentiates cognitive, social, vision, and biopsychological research at the second level, and identifies finer grained distinctions within these subjects at the leaves.

To extend the topic model to incorporate collocations, we introduced an additional set of variables that indicate whether a word is part of a collocation. Each word w_i thus has a topic assignment z_i and a collocation assignment x_i. The x_i variables can take on two values. If $x_i = 1$, then w_i is part of a collocation and is generated from a distribution that depends just on the previous word, $P(w_i|w_{i-1}, x_i = 1)$. If $x_i = 0$, then w_i is generated from the distribution associated with its topic, $P(w_i|z_i, x_i = 0)$. Importantly, the value of x_i is chosen based on the previous word, w_{i-1}, being drawn from the distribution $P(x_i|w_{i-1})$. This means that the model can capture dependencies between words: if w_{i-1} is UNITED, it is likely that $x_i = 1$, meaning that w_i is generated based just on the fact that it follows UNITED, and not on the topic. The graphical model corresponding to this extended generative process is shown in Figure 15. A more detailed description of the model appears in the Appendix C, together with an algorithm that can be used to simultaneously learn $P(w_i|w_{i-1}, x_i = 1)$, $P(w_i|z_i, x_i = 0)$, and $P(x_i = 1|w_{i-1})$ from a corpus.

Using this extended topic model, the conditional probability of one word given another is simply

$$P(w_2|w_1) = P(w_2|w_1, x_2 = 1)P(x_2 = 1|w_1) + P(w_2|w_1, x_2 = 0)P(x_2 = 0|w_1) \qquad (12)$$

where $P(w_2|w_1, x_2 = 0)$ is computed as in the basic topic model, using Equation 9. Thus, w_2 will be highly associated with w_1 either if w_2 tends to follow w_1, or if the two words tend to occur

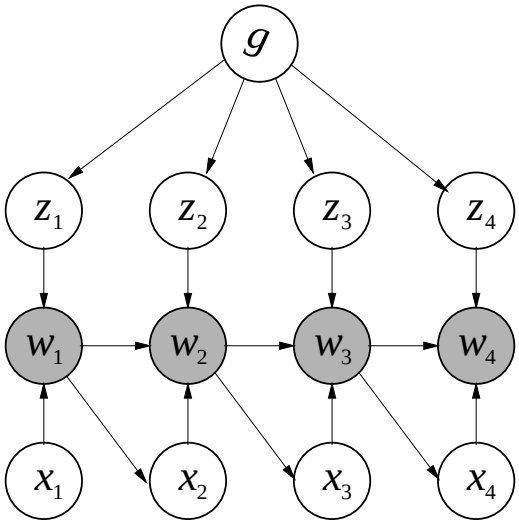

Figure 15. Graphical model indicating dependencies among variables in the collocation model. The variable x_i determines whether the word w_i is generated from a distribution that depends only on the previous word, being a collocation, or from a distribution that depends only on the topic z_i.

in the same semantic contexts. We used the algorithm described in Appendix C to estimate the probabilities required to compute $P(w_2|w_1)$ from the TASA corpus, using the same procedure to remove stop words as in our previous analyses, but supplying the words to the algorithm in the order that they actually occurred within each document. We then examined how well solutions with 500, 900, 1300, and 1700 topics predicted the word association norms collected by Nelson et al. (1998).

Introducing the capacity to produce collocations changes the associates that the model identifies. One way to see this is to examine cue-associate pairs produced by people that are in the set of ten words for which $P(w_2|w_1)$ is highest under the collocation model, but not in this set for the basic topic model. Considering just the first associates people produce and using the 1700 topic model, we find pairs such as UNITED-KINGDOM, BUMBLE-BEE, STORAGE-SPACE, METRIC-SYSTEM, MAIN-STREET, EVIL-DEVIL, FOREIGN-LANGUAGE, FRIED-CHICKEN, STOCK-MARKET, INTERSTATE-HIGHWAY, BOWLING-BALL, and SERIAL-NUMBER. These examples thus show how the collocation model is able to predict some associations that are based on word order rather than semantic context. Table 5 compares the median ranks of the associates under the ordering imposed by $P(w_2|w_1)$ for the collocation model and the basic topic model. The results show that the models perform very similarly: adding the capacity to capture associations based on word order does not result in a major improvement in the performance of the model. Hutchinson (2003) suggests that 11.6% of associations result from word order, which would lead us to expect some improvement in performance. The lack of improvement may be a consequence of the fact that incorporating the extra process for modeling collocations reduces the amount of data that is available for estimating topics, meaning that the model fails to capture some semantic associations.

Integrating topics and syntax

The model described in the previous section provides an extremely simple solution to the question of how topic models can be extended to capture word order, but our approach also supports

Table 5: Median ranks of the collocation model and basic topic model in predicting word association

	Associate				
Number of topics	1st	2nd	3rd	4th	5th
500	27 (29)	66 (70)	104 (106)	139 (141)	171 (175)
900	22 (22)	59 (57)	105 (101)	134 (131)	159 (159)
1300	20 (20)	58 (56)	105 (99)	131 (128)	160 (163)
1700	19 (18)	57 (54)	102 (100)	131 (130)	166 (164)

Note: Numbers in parentheses show the performance of the basic topic model without collocations.

more sophisticated solutions. Generative models can be used to overcome a major weakness of most statistical models of language – that they tend to model either syntax or semantics (although recent work provides some exceptions, including Dennis, 2004, and Jones and Mewhort, 2006). Many of the models used in computational linguistics, such as hidden Markov models and probabilistic context-free grammars (Charniak, 1993; Jurafsky & Martin, 2000; Manning & Shütze, 1999), generate words purely based on sequential dependencies among unobserved syntactic classes, not modeling the variation in content that occurs across documents, while topic models generate words in a way that is intended to capture the variation across documents, but ignores sequential dependencies. In cognitive science, methods such as distributional clustering (Redington, Chater, & Finch, 1998) are used to infer the syntactic classes of words, while methods such as LSA are used to analyze their meaning, and it is not clear how these different forms of statistical analysis should be combined.

Generative models can be used to define a model that captures both sequential dependencies and variation in content across contexts. This hybrid model illustrates the appealing modularity of generative models. Because a probabilistic language model specifies a probability distribution over words in a document in terms of components which are themselves probability distributions over words, different models are easily combined by mixing their predictions or embedding one inside the other. Griffiths, Steyvers, Blei, and Tenenbaum (2005; see also Griffiths & Steyvers, 2003) explored a composite generative model for language, in which one of the probability distributions over words used in defining a syntactic model was replaced with a semantic model. This allows the syntactic model to choose when to emit a semantically appropriate word, and the semantic model to choose which word to emit. The syntactic model used in this case was extremely simple, but this example serves to illustrate two points: that a simple model can discover categories of words that are defined both in terms of their syntactic roles and their semantic roles, and that defining a generative model that incorporates both of these factors is straightforward. A similar strategy could be pursued with a more complex probabilistic model of syntax, such as a probabilistic context-free grammar.

The structure of the composite generative model is shown in Figure 16. In this model, a word can appear in a document for two reasons: because it fulfills a functional syntactic role, or because it contributes to the semantic content. Accordingly, the model has two parts, one responsible for capturing sequential dependencies produced by syntax, and the other expresssing semantic dependencies. The syntactic dependencies are introduced via a hidden Markov model, a popular probabilistic model for language that is essentially a probabilistic regular grammar (Charniak, 1993; Jurafsky & Martin, 2000; Manning & Shütze, 1999). In a hidden Markov model, each word w_i is generated by first choosing a class c_i from a distribution that depends on the class of the previous word, c_{i-1}, and

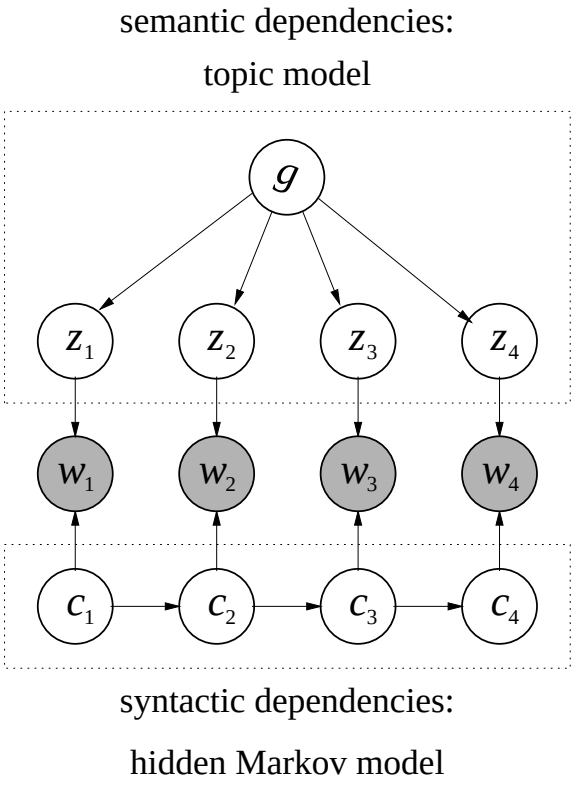

Figure 16. Graphical model indicating dependencies among variables in the composite model, in which syntactic dependencies are captured by a hidden Markov model (with the c_i variables being the classes from which words are generated) and semantic dependencies are captures by a topic model.

then generating w_i from a distribution that depends on c_i. The composite model simply replaces the distribution associated with one of the classes with a topic model, which captures the long-range semantic dependencies among words. An algorithm similar to that described in Appendix A can be used to infer the distributions over words associated with the topics and classes from a corpus (Griffiths et al., 2005).

The results of applying the composite model to a combination of the TASA and Brown (Kucera & Francis, 1967) corpora are shown in Figure 17. The factorization of words into those that appear as a result of syntactic dependencies (as represented by the class distributions) and those that appear as a result of semantic dependencies (represented by the topic distributions) pulls apart function and content words. In addition to learning a set of semantic topics, the model finds a set of syntactic classes of words that discriminate determiners, prepositions, pronouns, adjectives, and present- and past-tense verbs. The model performs about as well as a standard hidden Markov model – which is a state-of-the-art method – for identifying syntactic classes, and outperforms distributional clustering (Redington et al., 1998) in this task (Griffiths et al., 2005).

The ability to identify categories of words that capture their syntactic and semantic roles, based purely on their distributional properties, could be a valuable building block for the initial stages of language learning or for facilitating the extraction of gist. For example, learning the syntax of natural language requires a child to discover the rules of the grammar as well as the

Semantic topics

EARTH	KING	COLOR	FOOD	MONEY	FATHER	JOB	FARMERS
MOON	PRINCE	RED	STOMACH	GOLD	FAMILY	WORK	LAND
SUN	*	WHITE	MOUTH	DOLLARS	MOTHER	BUSINESS	CROPS
SPACE	GOLD	BLUE	TUBE	SILVER	CHILDREN	OFFICE	FARM
PLANET	PRINCESS	COLORS	DIGESTIVE	DOLLAR	BROTHER	JOBS	FOOD
PLANETS	PALACE	BROWN	INTESTINE	COINS	PARENTS	ACCOUNTING	PEOPLE
MARS	EMPEROR	GREEN	BODY	CENTS	HOUSE	EMPLOYEES	FARMING
ATMOSPHERE	QUEEN	BLACK	DIGESTION	PAPER	BROTHERS	WORKERS	WHEAT
SURFACE	DRAGON	*	AIR	CURRENCY	HOME	COMPANY	FARMS
ORBIT	CASTLE	YELLOW	SMALL	EXCHANGE	SON	INFORMATION	CORN

Syntactic classes

THE	IN	HE	*	BE	IS	CAN	SAID
A	FOR	IT	NEW	HAVE	WAS	WOULD	MADE
HIS	TO	YOU	OTHER	SEE	ARE	WILL	USED
THIS	ON	THEY	FIRST	MAKE	WERE	COULD	CAME
THEIR	WITH	I	SAME	DO	HAD	MAY	WENT
THESE	AT	SHE	GREAT	KNOW	HAVE	HAD	FOUND
YOUR	BY	WE	GOOD	GET	HAS	MUST	CALLED
HER	FROM	THERE	SMALL	GO		DO	
MY	AS	THIS	LITTLE	TAKE		HAVE	
SOME	INTO	WHO	OLD	FIND		DID	

Figure 17. Results of applying a composite model that has both syntactic and semantic latent structure to a concatenation of the TASA and Brown corpora. The model simultaneously finds the kind of semantic topics identified by the topic model and syntactic classes of the kind produced by a hidden Markov model.

abstract syntactic categories over which those rules are defined. These syntactic categories and rules are defined only with respect to each other, making it hard to see how one could learn both starting with neither. The syntactically organized word classes discovered by our simple statistical model could provide a valuable starting point for learning syntax, even though the notion of syntactic structure used in the model is far too simplistic to capture the syntax of English or any other natural language. The capacity to separate out the critical semantic content words in a document, from those words playing primarily syntactic functions, could also be valuable for modeling adult language processing or in machine information-retrieval applications. Only the semantic content words would be relevant, for example, in identifying the gist of a document or sentence. The syntactic function words can be – and usually are, by expert language processors – safely ignored.

Summary

Using generative models as a foundation for specifying psychological accounts of linguistic processing and semantic memory provides a way to define models that can be extended to incorporate more complex aspects of the structure of language. The extensions to the topic model described in this section begin to illustrate this potential. We hope to use this framework to develop statistical models that allow us to infer rich semantic structures that provide a closer match to the human semantic representation. In particular, the modularity of generative models provides the basis for exploring the interaction between syntax and semantics in human language processing, and suggests how different kinds of representation can be combined in solving computational problems that arise in other contexts.

Conclusion

Part of learning and using language is identifying the latent semantic structure responsible for generating a set of words. Probabilistic generative models provide solutions to this problem, making it possible to use powerful statistical learning to infer structured representations. The topic model

is one instance of this approach, and is a starting point for exploring how generative models can be used to address questions about human semantic representation. It outperforms Latent Semantic Analysis, a leading model of the acquisition of semantic knowledge, in predicting word association and a variety of other linguistic processing and memory tasks. It also explains several aspects of word association that are problematic for LSA: word frequency and asymmetry, violation of the triangle inequality, and the properties of semantic networks. The success of the model on these tasks comes from the structured representation that it assumes: by expressing the meaning of words in terms of different topics, the model is able to capture their different meanings and senses.

Going beyond the topic model, generative models provide a path towards a more comprehensive exploration of the role of structured representations and statistical learning in the acquisition and application of semantic knowledge. We have sketched some of the ways in which the topic model can be extended to bring it closer to the richness of human language. Although we are still far from understanding how people comprehend and acquire language, these examples illustrate how increasingly complex structures can be learned using statistical methods, and they show some of the potential for generative models to provide insight into the psychological questions raised by human linguistic abilities. Across many areas of cognition, perception, and action, probabilistic generative models have recently come to offer a unifying framework for understanding aspects of human intelligence as rational adaptations to the statistical structure of the environment (Anderson, 1990; Anderson & Schooler, 1991; Geisler et al., 2001; Griffiths & Tenenbaum, 2006b, 2006a; Kemp et al., 2004; Koerding & Wolpert, 2004; Simoncelli & Olshausen, 2001; Wolpert et al., 1995). It remains to be seen how far this approach can be carried in the study of semantic representation and language use, but the existence of large corpora of linguistic data and powerful statistical models for language clearly make this a direction worth pursuing.

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*, 261-295.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Bower, G. H. (1974). *Human associative memory.* Washington, DC: Hemisphere.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396-408.

Baldewein, U., & Keller, F. (2004). Modeling attachment decisions with a probabilistic parser: The case of head final structures. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Bartlett, F. C. (1932). *Remembering: a study in experimental and social psychology.* Cambridge: Cambridge University Press.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial processing tree modeling. *Psychonomic Bulletin and Review*, *6*, 57-86.

Bigi, B., De Mori, R., El-Beze, M., & Spriet, T. (1997). Combined models for topic spotting and topic-dependent language modeling. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings* (p. 535-542).

Blei, D., & Lafferty, J. (2006). Correlated topic models. In *Advances in neural information processing systems 18.* Cambridge, MA: MIT Press.

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems 16.* Cambridge, MA: MIT Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Brainerd, C. J., Reyna, V. F., & Mojardin, A. H. (1999). Conjoint recognition. *Psychological Review*, *106*, 160-179.

Brainerd, C. J., Wright, R., & Reyna, V. F. (2002). Dual-retrieval processes in free and associative recall. *Journal of Memory and Language*, *46*, 120-152.

Buntine, W. (2002). Variational extensions to EM and multinomial PCA. In *ECML 2002.*

Buntine, W., & Jakulin, A. (2004). Applying discrete PCA in data analysis. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI).* San Francisco, CA: Morgan Kaufmann.

Charniak, E. (1993). *Statistical language learning.* Cambridge, MA: MIT Press.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407-428.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behaviour*, *8*, 240-247.

Cramer, P. (1968). *Word association.* New York: Academic Press.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17-22.

Deese, J. (1962). On the structure of associative meaning. *Psychological Review*, *69*, 161-175.

Deese, J. (1965). *The structure of associations in language and thought.* Baltimore: Johns Hopkins University Press.

Dennis, S. (2003). A comparison of statistical models for the extraction of lexical information from text corpora. In *Proceedings of the Twenty-Fifth Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Dennis, S. (2004). An unsupervised method for the extraction of propositional information from text. *Proceedings of the National Academy of Sciences*, *101*, 5206-5213.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*, 452-478.

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, *27*, 429-446.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*, 211-245.

Erosheva, E. A. (2002). *Grade of membership and latent structure models with applications to disability survey data.* Unpublished doctoral dissertation, Department of Statistics, Carnegie Mellon University.

Ervin, S. M. (1961). Changes with age in the verbal determinants of word association. *American Journal of Psychology*, *74*, 361-372.

Fillenbaum, S., & Rapoport, A. (1971). *Structures in the subjective lexicon.* New York: Academic Press.

Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature, 368*, 542-545.

Galton, F. (1880). Psychometric experiments. *Brain, 2*, 149-162.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research, 41*, 711-724.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* New York: Chapman & Hall.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice.* Suffolk, UK: Chapman and Hall.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Griffiths, T. L., & Steyvers, M. (2003). Prediction and semantic association. In *Neural information processing systems 15*. Cambridge, MA: MIT Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science, 101*, 5228-5235.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17.* Cambridge, MA: MIT Press.

Griffiths, T. L., & Tenenbaum, J. B. (2006a). From mere coincidences to meaningful discoveries. *Cognition*. (Manuscript in press)

Griffiths, T. L., & Tenenbaum, J. B. (2006b). Optimal predictions in everyday cognition. *Psychological Science*. (Manuscript in press)

Hobbes, T. (1651/1998). *Leviathan.* Oxford: Oxford University Press.

Hofmann, T. (1999). Probablistic latent semantic indexing. In *Proceedings of the Twenty-Second Annual International SIGIR Conference.*

Hutchinson, K. A. (2003). Is semantic priming due to association strength or feature overlap. *Psychonomic Bulletin and Review, 10*, 785-813.

Iyer, R., & Ostendorf, M. (1996). Modeling long distance dependence in language: Topic mixtures vs. dynamic cache models. In *Proceedings of the International Conference on Spoken Language Processing 1* (p. 236-239).

James, W. (1890). *Principles of psychology.* New York: Holt.

Jones, M. N., & Mewhort, D. J. K. (2006). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*. (Manuscript in press)

Jordan, M. I. (1998). *Learning in graphical models.* Cambridge, MA: MIT Press.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20*, 137-194.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing.* Upper Saddle River, NJ: Prentice Hall.

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language, 32*, 474-516.

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective.* Cambridge, MA: Harvard University Press.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*, 163-182.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitkin, R. W. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies.* Edinburgh, UK: University Press.

Koerding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244-248.

Krumhansl, C. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, *85*, 450-463.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

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*, 211-240.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, *28*, 203-208.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*, 252-271.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

Markman, A. B. (1998). *Knowledge representation.* Hillsdale, NJ: Erlbaum.

Marr, D. (1982). *Vision.* San Francisco, CA: W. H. Freeman.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*, 295-323.

McEvoy, C. L., Nelson, D. L., & Komatsu, T. (1999). What's the connection between true and false memories: The different roles of inter-item associations in recall and recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*, 1177-1194.

McNeill, D. (1966). A study of word association. *Journal of Verbal Learning and Verbal Behavior*, *2*, 250-262.

Miller, G. A., & Fellbaum, C. (1998). *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

Minka, T., & Lafferty, J. (2002). Expectation-Propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI).* San Francisco, CA: Morgan Kaufmann.

Muller, P., & Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science*, *19*, 95-110.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*, 249-265.

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, *4*, 648-654.

Nelson, D. L., & Goodmon, L. B. (2002). Experiencing a word can prime its accessibility and its associative connections to related words. *Memory & Cognition*, 380-398.

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition*, *28*, 887-899.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The university of south florida word association, rhyme, and word fragment norms.* (http://www.usf.edu/FreeAssociation/)

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the infuence of impicitly activated memories on recall and recognition. *Psychological Review*, *105*, 299-324.

Nelson, D. L., Zhang, N., & McKinney, V. M. (2001). The ties that bind what is known to the recognition of what is new. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*, 1147-1159.

Norman, D. A., Rumelhart, D. E., & the LNR Research Group. (1975). *Explorations in cognition.* San Francisco, CA: W. H. Freeman.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, *23*, 94-140.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Francisco, CA: Morgan Kaufmann.

Pinker, S. (1999). *Words and rules: the ingredients of language.* New York: Basic books.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes*, *12*, 765-805.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, *48*, 21-69.

Potter, M. C. (1993). Very short term conceptual memory. *Memory & Cognition*, *21*, 156-161.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945-955.

Purver, M., Kording, K. P., Griffiths, T. L., & Tenenbaum, J. B. (2006). Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of COLING/ACL*.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*, 191-201.

Rayner, K., & Frazier, L. (1989). Seection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 779-790.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425-469.

Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, *25*, 337-354.

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, *28*, 89-104.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803-814.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin and Review*, *8*, 385-407.

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

Roget, P. (1911). *Roget's thesaurus of English words and phrases.* Available from Project Gutenberg.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the twentieth conference on uncertainty in artificial intelligence (uai)*. San Francisco, CA: Morgan Kaufmann.

Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). Cambridge, MA: MIT Press.

Sereno, S. C., O'Donnell, P. J., & Rayner, K. (2006). Eye movements and lexical ambiguity resolution: Investigating the subordinate-bias effect. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 335-350.

Sereno, S. C., Pacht, J. M., & Rayner, K. (1992). The effect of meaning frequency on processing lexically ambiguous words: Evidence form eye fixations. *Psychological Science*, *3*, 296-300.

Shepard, R., & Arabie, P. (1979). Additive clutering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*, 87-123.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Simoncelli, E. P., & Olshausen, B. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193-1216.

Steyvers, M., & Malmberg, K. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *29*, 760-766.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. In *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer.* Washington, DC: American Psychological Association.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *The tength acm sigkdd international conference on knowledge discovery and data mining.*

Steyvers, M., & Tenenbaum, J. B. (2005). The large scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*, 41-78.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2004). Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems 17.* Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629-641.

Till, R. E., Mross, E. F., & Kintsch, W. (1988). Time course of priming for associate and inference words in a discourse context. *Memory & Cognition*, *16*, 283-298.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327-352.

Tversky, A., & Gati, I. (1982). Similarity, separability and the triangle inequality. *Psychological Review*, *89*, 123-154.

Tversky, A., & Hutchinson, J. W. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, *93*, 3-22.

Ueda, N., & Saito, K. (2003). Parametric mixture models for multi-labeled text. In *Advances in Neural Information Processing Systems 15.* Cambridge: MIT Press.

Warrington, E. K. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology*, *27*, 635-657.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*, 440-442.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*, 598-604.

Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. (1998). Learning from text: Matching readers and text by latent semantic analysis. *Discourse Processes*, *25*, 309-336.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*, 1880-1882.

Yantis, S., Meyer, D. E., & Smith, J. E. K. (1991). Analyses of multinomial mixture distributions: New tests for stochastic models of cognition and action. *Psychological Bulletin*, *110*, 350-374.

Zipf, G. K. (1965). *Human behavior and the principle of least effort.* New York: Hafner.

Appendix A: Statistical formulation of the topic model

A number of approaches to statistical modeling of language have been based upon probabilistic topics. The notion that a topic can be represented as a probability distribution over words appears in several places in the natural language processing literature (e.g., Iyer & Ostendorf, 1996). Completely unsupervised methods for extracting sets of topics from large corpora were pioneered by Hofmann (1999), in his Probabilistic Latent Semantic Indexing method (also known as the aspect model). Blei, Ng, and Jordan (2003) extended this approach by introducing a prior on the distribution over topics, turning the model into a genuine generative model for collections of documents. Ueda and Saito (2003) explored a similar model, in which documents are balanced mixtures of a small set of topics. All of these approaches use a common representation, characterizing the content of words and documents in terms of probabilistic topics.

The statistical model underlying many of these approaches has also been applied to data other than text. Erosheva (2002) describes a model equivalent to a topic model, applied to disability data. The same model has been applied to data analysis in genetics (Pritchard, Stephens, & Donnelly, 2000). Topic models also make an appearance in the psychological literature on data analysis (Yantis, Meyer, & Smith, 1991). Buntine (2002) pointed out a formal correspondence between topic models and principal component analysis, providing a further connection to LSA.

A multi-document corpus can be expressed as a vector of words $\mathbf{w} = \{w_1, \ldots, w_n\}$, where each w_i belongs to some document d_i, as in a word-document co-occurrence matrix. Under the generative model introduced by Blei et al. (2003), the gist of each document, g, is encoded using a multinomial distribution over the T topics, with parameters $\theta^{(d)}$, so for a word in document d, $P(z|g) = \theta_z^{(d)}$. The zth topic is represented by a multinomial distribution over the W words in the vocabulary, with parameters $\phi^{(z)}$, so $P(w|z) = \phi_w^{(z)}$. We then take a symmetric Dirichlet(α) prior on $\theta^{(d)}$ for all documents, a symmetric Dirichlet(β) prior on $\phi^{(z)}$ for all topics. The complete statistical model can thus be written as

$$
\begin{aligned}
w_i \mid z_i, \phi^{(z_i)} &\sim \text{Discrete}(\phi^{(z_i)}) \\
\phi^{(z)} &\sim \text{Dirichlet}(\beta) \\
z_i \mid \theta^{(d_i)} &\sim \text{Discrete}(\theta^{(d_i)}) \\
\theta^{(d)} &\sim \text{Dirichlet}(\alpha)
\end{aligned}
$$

The user of the algorithm can specify α and β, which are hyperparameters that affect the granularity of the topics discovered by the model (see Griffiths & Steyvers, 2004).

An algorithm for finding topics

Several algorithms have been proposed for learning topics, including expectation-maximization (EM; Hofmann, 1999), variational EM (Blei et al., 2003; Buntine, 2002), expectation propagation (Minka & Lafferty, 2002), and several forms of Markov chain Monte Carlo (MCMC; Buntine & Jakulin, 2004; Erosheva, 2002; Griffiths & Steyvers, 2002; 2003; 2004; Pritchard et al., 2000). We use Gibbs sampling, a form of Markov chain Monte Carlo.

We extract a set of topics from a collection of documents in a completely unsupervised fashion, using Bayesian inference. The Dirichlet priors are conjugate to the multinomial distributions ϕ, θ, allowing us to compute the joint distribution $P(\mathbf{w}, \mathbf{z})$ by integrating out ϕ and θ. Since

$P(\mathbf{w}, \mathbf{z}) = P(\mathbf{w}|\mathbf{z})P(\mathbf{z})$ and ϕ and θ only appear in the first and second terms respectively, we can perform these integrals separately. Integrating out ϕ gives the first term

$$P(\mathbf{w}|\mathbf{z}) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{j=1}^{T} \frac{\prod_w \Gamma(n_j^{(w)} + \beta)}{\Gamma(n_j^{(\cdot)} + W\beta)}, \tag{13}$$

in which $n_j^{(w)}$ is the number of times word w has been assigned to topic j in the vector of assignments $\mathbf{z}$ and $\Gamma(\cdot)$ is the standard gamma function. The second term results from integrating out θ, to give

$$P(\mathbf{z}) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{d=1}^{D} \frac{\prod_j \Gamma(n_j^{(d)} + \alpha)}{\Gamma(n_\cdot^{(d)} + T\alpha)},$$

where $n_j^{(d)}$ is the number of times a word from document d has been assigned to topic j. We can then ask questions about the posterior distribution over $\mathbf{z}$ given $\mathbf{w}$, given by Bayes rule:

$$P(\mathbf{z}|\mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{z})}{\sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z})}.$$

Unfortunately, the sum in the denominator is intractable, having T^n terms, and we are forced to evaluate this posterior using Markov chain Monte Carlo.

Markov chain Monte Carlo (MCMC) is a procedure for obtaining samples from complicated probability distributions, allowing a Markov chain to converge to the target distribution and then drawing samples from the Markov chain (see Gilks, Richardson & Spiegelhalter, 1996). Each state of the chain is an assignment of values to the variables being sampled, and transitions between states follow a simple rule. We use Gibbs sampling, where the next state is reached by sequentially sampling all variables from their distribution when conditioned on the current values of all other variables and the data. We sample only the assignments of words to topics, z_i.

The conditional posterior distribution for z_i is given by

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}, \tag{14}$$

where $\mathbf{z}_{-i}$ is the assignment of all z_k such that $k \neq i$, and $n_{-i,j}^{(w_i)}$ is the number of words assigned to topic j that are the same as w_i, $n_{-i,j}^{(\cdot)}$ is the total number of words assigned to topic j, $n_{-i,j}^{(d_i)}$ is the number of words from document d_i assigned to topic j, and $n_{-i,\cdot}^{(d_i)}$ is the total number of words in document d_i, all not counting the assignment of the current word w_i.

The MCMC algorithm is then straightforward. The z_i are initialized to values between 1 and T, determining the initial state of the Markov chain. The chain is then run for a number of iterations, each time finding a new state by running sampling each z_i from the distribution specified by Equation 14. After enough iterations for the chain to approach the target distribution, the current values of the z_i are recorded. Subsequent samples are taken after an appropriate lag, to ensure that their autocorrelation is low. Further details of the algorithm are provided in Griffiths and Steyvers (2004), where we show how it can be used to analyze the content of document collections.

The variables involved in the MCMC algorithm, and their modification across samples, are illustrated in Figure 18, which uses the data from Figure 2. Each word token in the corpus, w_i,

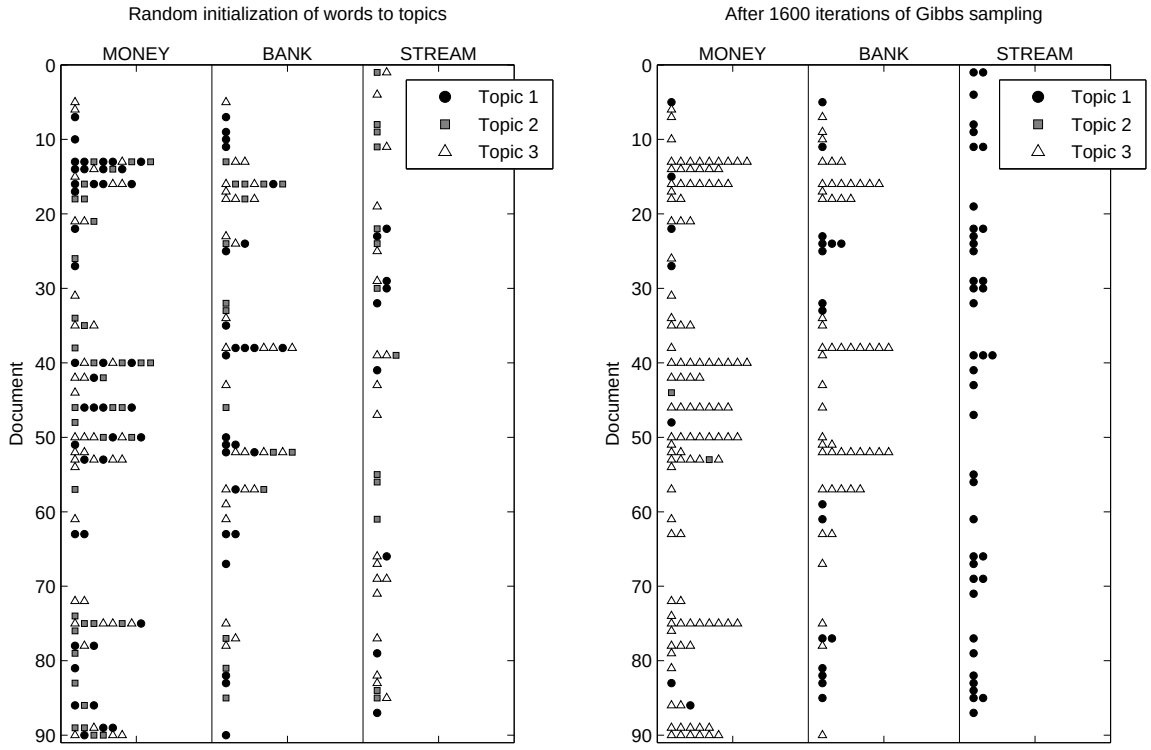

Figure 18. Illustration of the Gibbs sampling algorithm for learning topics, using the data from Figure 2. Each word token w_i appearing in the corpus has a topic assignment, z_i. The figure shows the assignments of all tokens of three types – MONEY, BANK, and STREAM – before and after running the algorithm. Each marker corresponds to a single token appearing in a particular document, and shape and color indicates assignment: topic 1 is a black circle, topic 2 is a gray square, and topic 3 is a white triangle. Before running the algorithm, assignments are relatively random, as shown in the left panel. After running the algorithm, tokens of MONEY are almost exclusively assigned to topic 3, tokens of STREAM are almost exclusively assigned to topic 1, and tokens of BANK are assigned to whichever of topic 1 and topic 3 seems to dominate a given document. The algorithm consists of iteratively choosing an assignment for each token, using a probability distribution over tokens that guarantees convergence to the posterior distribution over assignments.

has a topic assignment, z_i, at each iteration of the sampling procedure. In this case, we have 90 documents and a total of 731 words w_i, each with their own z_i. In the figure, we focus on the tokens of three words: MONEY, BANK, and STREAM. Each word token is initially randomly assigned to a topic, and each iteration of MCMC results in a new set of assignments of tokens to topics. After a few iterations, the topic assignments begin to reflect the different usage patterns of MONEY and STREAM, with tokens of these words ending up in different topics, and the multiple senses of BANK.

The result of the MCMC algorithm is a set of samples from $P(\mathbf{z}|\mathbf{w})$, reflecting the posterior distribution over topic assignments given a collection of documents. From any single sample we

can obtain an estimate of the parameters ϕ and θ from $\mathbf{z}$ via

$$\hat{\phi}_w^{(j)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + W\beta} \tag{15}$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha}. \tag{16}$$

These values correspond to the predictive distributions over new words w and new topics z conditioned on $\mathbf{w}$ and $\mathbf{z}$, and the posterior means of θ and ϕ given $\mathbf{w}$ and $\mathbf{z}$.

Prediction, disambiguation, and gist extraction

The generative model allows documents to contain multiple topics, which is important when modeling long and complex documents. Assume we have an estimate of the topic parameters, ϕ. Then the problems of prediction, disambiguation, and gist extraction can be reduced to computing

$$P(w_{n+1}|\mathbf{w};\phi) = \sum_{\mathbf{z},z_{n+1}} P(w_{n+1}|z_{n+1};\phi)P(z_{n+1}|\mathbf{z})P(\mathbf{z}|\mathbf{w};\phi) \tag{17}$$

$$P(\mathbf{z}|\mathbf{w};\phi) = \frac{P(\mathbf{w},\mathbf{z}|\phi)}{\sum_{\mathbf{z}} P(\mathbf{w},\mathbf{z}|\phi)} \tag{18}$$

$$P(g|\mathbf{w};\phi) = \sum_{\mathbf{z}} P(g|\mathbf{z})P(\mathbf{z}|\mathbf{w};\phi) \tag{19}$$

respectively. The sums over $\mathbf{z}$ that appear in each of these expressions quickly become intractable, being over T^n terms, but they can be approximated using MCMC.

In many situations, such as processing a single sentence, it is reasonable to assume that we are dealing with words that are drawn from a single topic. Under this assumption, g is represented by a multinomial distribution θ that puts all of its mass on a single topic, z, and $z_i = z$ for all i. The problems of disambiguation and gist extraction thus reduce to inferring z. Applying Bayes' rule,

$$\begin{aligned} P(z|\mathbf{w};\phi) &= \frac{P(\mathbf{w}|z;\phi)P(z)}{\sum_{\mathbf{z}} P(\mathbf{w}|z;\phi)P(z)} \\ &= \frac{\prod_{i=1}^n P(w_i|z;\phi)P(z)}{\sum_z \prod_{i=1}^n P(w_i|z;\phi)P(z)} \\ &= \frac{\prod_{i=1}^n \phi_{w_i}^{(z)}}{\sum_z \prod_{i=1}^n \phi_{w_i}^{(z)}}, \end{aligned}$$

where the last line assumes a uniform prior, $P(z) = \frac{1}{T}$, consistent with the symmetric Dirichlet priors assumed above. We can then form predictions via

$$\begin{aligned} P(w_{n+1}|\mathbf{w};\phi) &= \sum_z P(w_{n+1}, z|\mathbf{w};\phi) \\ &= \sum_z P(w_{n+1}|z;\phi)P(z|\mathbf{w};\phi) \\ &= \frac{\sum_z \prod_{i=1}^{n+1} \phi_{w_i}^{(z)}}{\sum_z \prod_{i=1}^n \phi_{w_i}^{(z)}} \end{aligned}$$

This predictive distribution can be averaged over the estimates of ϕ yielded by a set of samples from the MCMC algorithm.

For the results described in the paper, we ran three Markov chains for 1600 iterations at each value of T, using $\alpha = 50/T$ and $\beta = 0.01$. We started sampling after 800 iterations, taking one sample every 100 iterations thereafter. This gave a total of 24 samples for each choice of dimensionality. The topics shown in Table 7 are taken from a single sample from the Markov chain for the 1700 dimensional model. We computed an estimate of ϕ using Equation 15 and used these values to compute $P(w_2|w_1)$ for each sample, then averaged the results across all of the samples to get an estimate of the full posterior predictive distribution. This averaged distribution was used in evaluating the model on the word association data.

Appendix B: Topics and features

Tversky (1977) considered a number of different models for the similarity between two stimuli, based upon the idea of combining common and distinctive features. Most famous is the contrast model, in which the similarity between X and Y, $S(X, Y)$, is given by

$$S(X, Y) = \theta f(\mathcal{X} \cap \mathcal{Y}) - \alpha f(\mathcal{Y} - \mathcal{X}) - \beta f(\mathcal{X} - \mathcal{Y}),$$

where $\mathcal{X}$ is the set of features to which X belongs, $\mathcal{Y}$ is the set of features to which Y belongs, $X \cap Y$ is the set of common features, $\mathcal{Y} - \mathcal{X}$ is the set of distinctive features of Y, $f(\cdot)$ is a measure over those sets, and θ, α, β are parameters of the model. Another model considered by Tversky, which is also consistent with the axioms used to derive the contrast model, is the ratio model, in which

$$S(X, Y) = 1/\left[1 + \frac{\alpha f(\mathcal{Y} - \mathcal{X}) + \beta f(\mathcal{X} - \mathcal{Y})}{\theta f(\mathcal{X} \cap \mathcal{Y})}\right].$$

As in the contrast model, common features increase similarity and distinctive features decrease similarity. The only difference between the two models is the form of the function by which they are combined.

Tversky's (1977) analysis assumes that the features of X and Y are known. However, in some circumstances, possession of a particular feature may be uncertain. For some hypothetical feature h, we might just have a probability that X possesses h, $P(X \in h)$. One means of dealing with this uncertainty is replacing $f(\cdot)$ with its expectation with respect to the probabilities of feature possession. If we assume that $f(\cdot)$ is linear (as in additive clustering models, e.g., Shepard & Arabie, 1979) and gives uniform weight to all features, the ratio model becomes

$$S(X, Y) = 1/\left[1 + \frac{\alpha \sum_h (1 - P(X \in h))P(Y \in h) + \beta \sum_h (1 - P(Y \in h))P(X \in h)}{\theta \sum_h P(X \in h)P(Y \in h)}\right]. \quad (20)$$

where we take $P(X \in h)$ to be independent for all X and h. The sums in this Equation reduce to counts of the common and distinctive features if the probabilities all take on values of 0 or 1.

In the topic model, semantic association is assessed in terms of the conditional probability $P(w_2|w_1)$. This quantity reduces to

$$\begin{aligned} P(w_2|w_1) &= \frac{\sum_z P(w_2|z)P(w_1|z)}{\sum_z P(w_1|z)} \\ &= \frac{\sum_z P(w_2|z)P(w_1|z)}{\sum_z P(w_2|z)P(w_1|z) + \sum_z (1 - P(w_2|z))P(w_1|z)} \\ &= 1/\left[1 + \frac{\sum_z (1 - P(w_2|z))P(w_1|z)}{\sum_z P(w_2|z)P(w_1|z)}\right], \end{aligned}$$

which can be seen to be of the same form as the probabilistic ratio model specified in Equation 20, with $\alpha = 1$, $\beta = 0$, $\theta = 1$, topics z in the place of features h, and $P(w|z)$ replacing $P(X \in h)$. This result is similar to that of Tenenbaum and Griffiths (2001), who showed that their Bayesian model of generalization was equivalent to the ratio model.

Appendix C: The collocation model

Using the notation introduced above, the collocation model can be written as

$$\begin{aligned}
w_i \mid z_i, x_i = 0, \phi^{(z_i)} &\sim \text{Discrete}(\phi^{(z_i)}) \\
w_i \mid w_{i-1}, x_i = 1, \phi^{(w_{i-1})} &\sim \text{Discrete}(\phi^{(w_{i-1})}) \\
\phi^{(z)} &\sim \text{Dirichlet}(\beta) \\
\phi^{(w)} &\sim \text{Dirichlet}(\delta) \\
z_i \mid \theta^{(d_i)} &\sim \text{Discrete}(\theta^{(d_i)}) \\
\theta^{(d)} &\sim \text{Dirichlet}(\alpha) \\
x_i \mid w_{i-1} &\sim \text{Discrete}(\pi^{(w_{i-1})}) \\
\pi^{(w)} &\sim \text{Beta}(\gamma_0, \gamma_1)
\end{aligned}$$

where $\phi^{(w_i)}$ is the distribution over w_i given w_{i-1}, and $\pi^{(w_{i-1})}$ is the distribution over x_i given w_{i-1}. The Gibbs sampler for this model is as follows. If $x_i = 0$, then z_i is drawn from the distribution

$$P(z_i|\mathbf{z}_{-i}, \mathbf{w}, \mathbf{c}) \propto \frac{n_{w_i}^{(z_i)} + \beta}{n_\cdot^{(z_i)} + W\beta} \frac{n_{z_i} + \alpha}{n + T\alpha} \tag{21}$$

where all counts exclude the current case and only refer to the words for which $x_i = 0$, which are the words assigned to the topic model (e.g., n is the total number of words for which $x_i = 0$, not the total number of words in the corpus). If $x_i = 1$, then z_i is sampled from

$$P(z_i|\mathbf{z}_{-i}, \mathbf{w}, \mathbf{c}) \propto \frac{n_{z_i} + \alpha}{n + T\alpha} \tag{22}$$

where again the counts are only for the words for which $x_i = 0$. Finally, x_i is drawn from the distribution

$$P(x_i|\mathbf{x}_{-i}, \mathbf{w}, \mathbf{z}) \propto \begin{cases} \frac{n_{w_i}^{(z_i)}+\beta}{n_\cdot^{(z_i)}+W\beta} \frac{n_0^{(w_{i-1})}+\gamma_0}{n_\cdot^{(w_{i-1})}+\gamma_0+\gamma_1} & x_i = 0 \\ \frac{n_{w_i}^{(w_{i-1})}+\delta}{n_\cdot^{(w_{i-1})}+W\delta} \frac{n_1^{(w_{i-1})}+\gamma_1}{n_\cdot^{(w_{i-1})}+\gamma_0+\gamma_1} & x_i = 1 \end{cases} \tag{23}$$

where $n_0^{(w_{i-1})}$ and $n_1^{(w_{i-1})}$ are the number of times the word w_{i-1} has been drawn from a topic or formed part of a collocation resectively, and all counts exclude the current case.

To estimate the parameters of the model for each sample, we can again use the posterior mean. The estimator for $\phi^{(z)}$ is just $\hat{\phi}^{(z)}$ from Equation 15. A similar estimator exists for the distribution associated with successive words

$$\hat{\phi}_{w_2}^{(w_1)} = \frac{n_{w_2}^{(w_1)} + \delta}{n_\cdot^{(w_1)} + W\delta}. \tag{24}$$

For $\pi^{(w_1)}$, which is the estimate of the probability that $x_2 = 1$ given w_1, we have

$$\hat{pi}^{(w_1)} = \frac{n_1^{(w_1)} + \gamma_1}{n_\cdot^{(w_1)} + \gamma_0 + \gamma_1}. \tag{25}$$

Using these estimates, Equation 12 becomes

$$P(w_2|w_1) = \pi^{(w_1)}\phi_{w_2}^{(w_1)} + (1 - \pi^{(w_1)}) \sum_z \phi_{w_2}^{(z)} \frac{\phi_{w_1}^{(z)}}{\sum_z \phi_{w_1}^{(z)}}. \quad (26)$$

The results described in the paper were averaged over 24 samples produced by the MCMC algorithm, with $\beta = 0.01$, $\alpha = 50/T$, $\gamma_0 = 0.1$, $\gamma_1 = 0.1$ and $\delta = 0.1$. The samples were collected from three chains in the same way as for the basic topic model.