Topics in Cognitive Science 2 (2010) 101–113

ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2009.01074.x

The Hidden Markov Topic Model: A Probabilistic Model of Semantic Representation

Mark Andrews, Gabriella Vigliocco

Department of Cognitive, Perceptual, and Brain Sciences, Division of Psychology and Language Sciences, University College London

Received 2 October 2009; accepted 12 October 2009

Abstract

In this paper, we describe a model that learns semantic representations from the distributional statistics of language. This model, however, goes beyond the common *bag-of-words* paradigm, and infers semantic representations by taking into account the inherent sequential nature of linguistic data. The model we describe, which we refer to as a *Hidden Markov Topics model*, is a natural extension of the current state of the art in Bayesian bag-of-words models, that is, the Topics model of Griffiths, Steyvers, and Tenenbaum (2007), preserving its strengths while extending its scope to incorporate more fine-grained linguistic information.

Keywords: Bayesian models; Probabilistic models; Computational models; Semantic representation; Semantic memory

1. Introduction

How word meanings are represented and learned is a foundational problem in the study of human language use. Within cognitive science, a promising recent approach to this problem has been the study of how the meanings of words can be learned from their statistical distribution across the language. This approach is motivated by the so-called *distributional hypothesis*, originally due to Harris (1954) and Firth (1957), which proposes that the meaning of a word can be derived from the linguistic contexts in which it occurs. Numerous large-scale computational implementations of this approach—including, for example, the

Correspondence should be sent to Mark Andrews, Department of Cognitive, Perceptual, and Brain Sciences, Division of Psychology and Language Sciences, University College London, 2 Bedford Way, London WC1H OAP. E-mail: m.andrews@ucl.ac.uk

work of Schüutze (1992), the HAL model (e.g., Lund, Burgess, & Atchley, 1995), the LSA model (e.g., Landauer & Dumais, 1997) and, most recently, the Topics model (e.g., Griffiths, Steyvers, & Tenenbaum, 2007)—have successfully demonstrated that the meanings of words can, at least in part, be derived from their statistical distribution in language.

Important as these computational models have been, one of their widely shared practices has been to treat the linguistic contexts in which a word occurs as unordered sets of words. In other words, the linguistic context of any given word is defined by which words co-occur with it and with what frequency, but it disregards all fine-grained sequential and syntactic information. By disregarding these types of data, these so-called *bag-of-words* models drastically restrict the information from which word meanings can be learned. All languages have strong syntactic-semantic correlations. The sequential order in which words occur, the argument structure and general syntactic relationships within sentences, all provide vital information about the possible meaning of words. This information is unavailable in bag-of-words models and consequently the extent to which they can extract semantic information from text, or adequately model human semantic learning, is limited.

In this paper, we describe a distributional model that goes beyond the bag-of-words paradigm. This model is a natural extension to the current state of the art in probabilistic bag-of-words models, namely the Topics model described in Griffiths et al. (2007) and elsewhere. The model we propose is a seamless continuation of the Topics model, preserving its strengths—its thoroughly unsupervised learning, its hierarchical Bayesian nature—while extending its scope to incorporate more fine-grained sequential and syntactic data.

2. The Topics model

The standard Topics model as described in Griffiths and Steyvers (2002, 2003); Griffiths et al. (2007) is a probabilistic generative model for texts and is based on the latent Dirichlet allocation (LDA) model of Blei, Ng, and Jordan (2003). It stipulates that each word in a corpus of texts is drawn from one of K latent distributions $\phi_1 \ldots \phi_k \ldots \phi_K \stackrel{\circ}{=} \boldsymbol{\phi}$, with each ϕ_k being a probability distribution over the V word-types in a fixed vocabulary. These distributions are the so-called *topics* that give the model its name. Some examples, learned by a Topics model described in Andrews, Vigliocco, and Vinson (2009), are given in Table 1. As is evident from this table, each topic is a cluster of related terms that corresponds to a

Table 1
Topics taken from an LDA model described in Andrews et al. (2009)

theatre	music	league	prison	rate	pub	market	railway	air
stage	band	cup	years	cent	guinness	stock	train	aircraft
arts	rock	season	sentence	inflation	beer	exchange	station	flying
play	song	team	jail	recession	drink	demand	steam	flight
dance	record	game	home	recovery	bar	share	rail	plane
opera	pop	match	prisoner	economy	drinking	group	engine	airport
cast	dance	division	serving	cut	alcohol	news	track	pilot

Each column gives the seven most probable word types in each topic.

coherent semantic theme, or subject-matter. As such, the topic distributions correspond to the semantic knowledge learned by the model, and the semantic representation of each word in the vocabulary is given by a distribution over them.

To describe the Topics model more precisely, let us assume we have a corpus of J texts $\mathbf{w}_1 \ldots \mathbf{w}_j \ldots \mathbf{w}_J$, where the jth text $\mathbf{w}_j$ is a sequence of n_j words, that is, $w_{j1} \ldots w_{ji} \ldots w_{jn_j}$. Each word, in each text, is assumed to be sampled from one of the model's K topics. Which one of these topics is chosen is determined by the value of a latent variable x_{ji} that corresponds to each word w_{ji}. This variable takes on one of K discrete values and is determined by sampling from a probability distribution π_j, which is specific to text j. As such, we can see that each text $\mathbf{w}_j$ is assumed to be a set of n_j independent samples from a mixture model. This mixture model is specific to text j, as the mixing distribution is determined by π_j. However, across texts, all mixing distributions are drawn from a common Dirichlet distribution, with parameters $\boldsymbol{\alpha}$. Given known values for $\boldsymbol{\phi}$ and $\boldsymbol{\alpha}$, the likelihood of the entire corpus is given by

$$\mathrm{P}(\mathbf{w}_{1:J}|\boldsymbol{\phi}, \boldsymbol{\alpha}) = \prod_{j=1}^{J} \int \mathrm{d}\pi_j \mathrm{P}(\pi_j|\boldsymbol{\alpha}) \left[\prod_{i=1}^{n_j} \sum_{\{x_{ji}\}} \mathrm{P}(w_{ji}|x_{ji}, \boldsymbol{\phi}) \mathrm{P}(x_{ji}|\pi_j) \right], \tag{1}$$

In this, we see that the Dirichlet distribution $\mathrm{P}(\pi_j \mid \boldsymbol{\alpha})$ introduces a hierarchical coupling of all the mixing distributions. As such, the standard Topics model is a hierarchical coupling of mixture models, effectively defining a continuum of mixture models, all sharing the same K component topics.

The standard Topics model is thus an example of a *hierarchical language model*. Its Bayesian network is shown in Fig.1. It assumes that each text is generated according to an

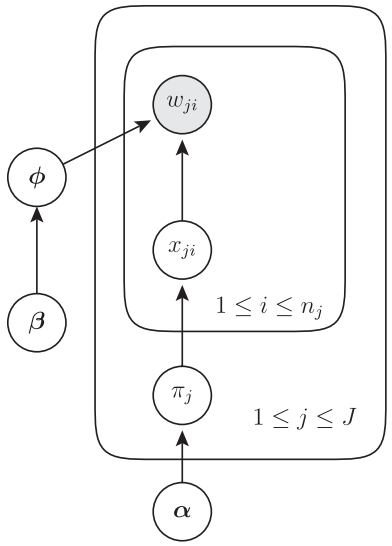


Fig. 1. A Bayesian network diagram of the standard Topics model described in Griffiths et al. (2007) and elsewhere. Details are provided in the main text. Note that $\boldsymbol{\beta}$ denotes the parameters of a V-dimensional Dirichlet distribution, from which each of K topic distributions is sampled.

elementary language model—specifically a mixture of unigram distributions—which are then hierarchically coupled with one another. From this, it is evident that the standard model can be extended by simply changing the elementary language model on which it is based. There are multiple possible language models that could be used in this respect. One possibility is to use a bigram language model as described in Wallach (2006). Another possibility is to use a language model based on a full phrase-structure grammar as described in Boyd-Graber and Blei (2009). However, in that work, the syntactic structure underlying the sentences in the texts is assumed to be known in advance and is provided by syntactically tagged corpus. In what follows, we describe a Topics models whose elementary language model is a Hidden Markov model (HMM). We refer to this as the *Hidden Markov Topics model* (HMTM).

3. Hidden Markov Topics model

In the HMTM, just as in the standard Topics model, each w_{ji} is drawn from one of K topics, the identity of which is determined by the latent variable x_{ji}. However, rather than sampling each x_{ji} independently from a probability distribution π_j, as in the standard model, these latent variables are generated by a Markov chain that is specific to text j. By direct analogy with the standard model, across the texts, the parameters of these Markov chains are drawn from a common set of Dirichlet distributions. As such, the HMTM is a hierarchical coupling of HMMs, defining a continuum of Hidden Markov models, all sharing the same state to output mapping.

In the HMTM, the likelihood of the corpus is

$$\mathrm{P}(\mathbf{w}_{1:J}|\boldsymbol{\phi},\boldsymbol{\alpha},\boldsymbol{\gamma}) = \prod_{j=1}^{J} \int \mathrm{d}\pi_j \mathrm{d}\theta_j \mathrm{P}(\pi_j,\theta_j|\boldsymbol{\alpha},\boldsymbol{\gamma}) \left[\sum_{\{\mathbf{x}_j\}} \prod_{i=1}^{n_j} \mathrm{P}(w_{ji}|x_{ji},\boldsymbol{\phi}) \prod_{i=2}^{n_j} \mathrm{P}(x_{ji}|x_{ji-1},\boldsymbol{\theta}_j)\mathrm{P}(x_{j1}|\boldsymbol{\pi}_j) \right]. \quad (2)$$

Here, θ_j and π_j are the parameters of the Markov chain of latent-states in text j. The θ_j is the $K \times K$ state transition matrix (i.e., the kth row of θ_j gives the probability of transitioning to each of the K states, given that the current state is k), and π_j is the initial distribution for the K states. The distribution over the π_j and θ_j, that is, $\mathrm{P}(\pi_j,\theta_j \mid \boldsymbol{\alpha},\gamma)$, is a set of independent Dirichlet distributions, where $\boldsymbol{\alpha}$ are the parameters for the Dirichlet distribution over π_j, and $\boldsymbol{\gamma}_1 \ldots \boldsymbol{\gamma}_K \overset{\circ}{=} \boldsymbol{\gamma}$ are the parameters for the distribution over each of the K rows of θ_j.

3.1. Learning and inference

From this description of the HMTM, as well as from its Bayesian network diagram given in Fig. 2, we can see that the HMTM has four sets of parameters whose values must be inferred from a training corpus of texts. These are the K topic distributions, that is, $\boldsymbol{\phi}$, the Dirichlet parameters for the latent-variable Markov chains, that is, $\boldsymbol{\alpha}$ and γ, and the Dirichlet

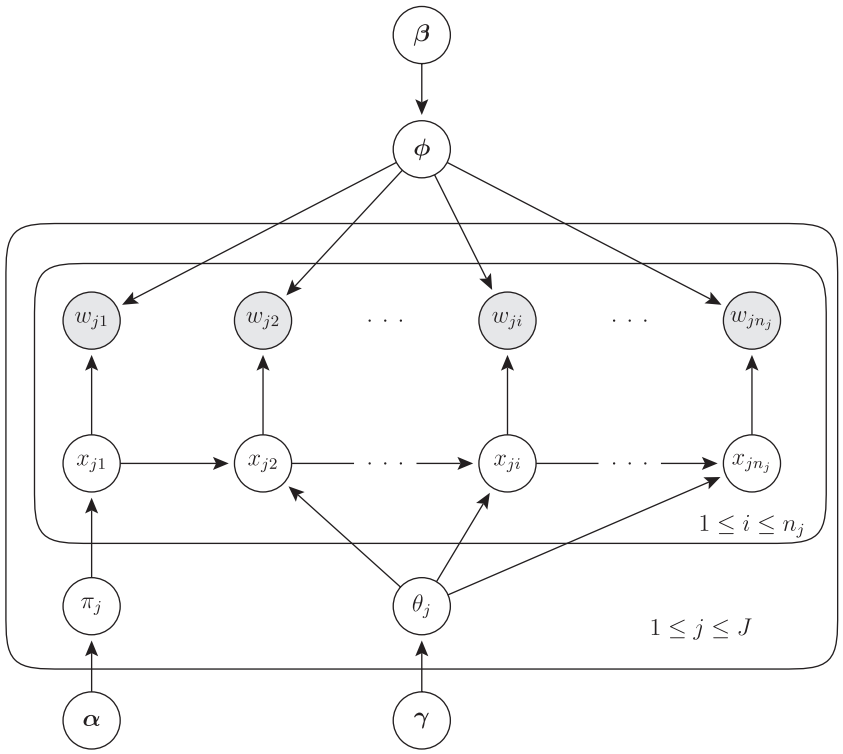


Fig. 2. A Bayesian network diagram for the Hidden Markov Topic model. As in the standard model, depicted in Fig. 1, each observed variable w_{ji} is assumed to be sampled from one of K topic distributions, collectively specified by $\boldsymbol{\phi}$. Likewise, which one of these distributions is chosen is determined by the value of the unobserved variable x_{ji}. In the standard model, however, the latent variables $x_{j1}\ldots x_{jn_j}$ are drawn independently from a distribution π_j, which is specific to text j. In the HMTM by contrast, $x_{j1}\ldots x_{jn_j}$ are drawn from a Markov chain specific to the jth text. The parameters of this chain are θ_j (i.e., the $K \times K$ state-space mapping for chain j) and π_j (i.e., the initial conditions for the jth chain). For all J texts, the parameters θ_j and π_j are drawn from a common set of Dirichlet distributions. Each θ_j is drawn from K-dimensional Dirichlet distributions, whose parameters are collectively specified by γ. Each π_j is drawn from a K-dimensional Dirichlet distribution $\boldsymbol{\alpha}$. Also shown here is $\boldsymbol{\beta}$, the parameters of a V-dimensional Dirichlet distribution from which the K topic distributions $\boldsymbol{\phi}$ are sampled.

parameters from which the topic distributions are drawn, that is, $\boldsymbol{\beta}$.[1] The posterior over $\boldsymbol{\phi}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and γ given a set of J texts $\mathbf{w}_{1:J}$ is

$$\mathrm{P}(\boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{w}_{1:J}) \propto \mathrm{P}(\mathbf{w}_{1:J} | \boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \mathrm{P}(\boldsymbol{\phi} | \boldsymbol{\beta}) \mathrm{P}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}), \tag{3}$$

where the likelihood term $\mathrm{P}(\mathbf{w}_{1:J} \mid \boldsymbol{\phi},\boldsymbol{\alpha},\gamma)$ is given by Eq. 2. This distribution is intractable (as is even the likelihood term itself), and as such it is necessary to use Markov Chain Monte Carlo (MCMC) methods to sample from it.

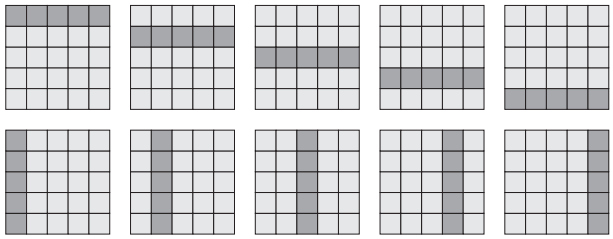
There are different options available for MCMC sampling. The method we employ is a Gibbs sampler that draws samples from the posterior over $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, γ, and $\mathbf{x}_{1:J}$. Here, $\mathbf{x}_{1:J}$ are the sequences of latent-variables for each of the J texts, that is, $\mathbf{x}_1$,$\mathbf{x}_2\ldots\mathbf{x}_J$, with $\mathbf{x}_j = x_{j1}\ldots x_{jn_j}$. This Gibbs sampler has the useful property of integrating over $\boldsymbol{\phi}$, $\boldsymbol{\pi}_{1:J}$ and $\boldsymbol{\theta}_{1:J}$, which entails both computational efficiency and faster convergence.

On convergence of the Gibbs sampler, we will obtain samples from the joint posterior distribution $P(\mathbf{x}_{1:J},\boldsymbol{\alpha},\boldsymbol{\beta},\gamma \mid \mathbf{w}_{1:J})$. From this, other variables of interest can be obtained. For example, it is desirable to know the likely values of the topic distributions $\phi_1 \ldots \phi_k \ldots \phi_K$. Given known values for $\mathbf{x}_{1:J}$ and $\boldsymbol{\beta}$, the posterior distribution over $\boldsymbol{\phi}$ is simply a product of Dirichlet distribution, from which samples are easily drawn and averaged. Further details of this Gibbs sampler are provided in the Appendix.

3.2. Demonstration

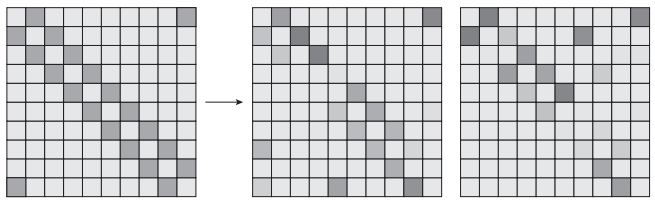
Here, we demonstrate the operation of the HMTM on a toy problem. In this problem, we generate a data-set from a HMTM with known values for $\boldsymbol{\phi}$, $\boldsymbol{\alpha}$, and γ. We can then train another HMTM, whose parameters are unknown, with this training data-set. Using the Gibbs sampler, we can draw samples from the posterior over $\boldsymbol{\phi}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and γ, and then compare these samples to the true parameter values that generated the training data.

In the example we present here, we use a ''vocabulary'' of $V = 25$ symbols and a set of $K = 10$ ''topics.'' As is common practice in demonstrations of related models, the ''topics'' we use in this example can be visualized using the grids shown below.



Each grid, although shown as a 5×5 square, is just a 25 dimensional array, with 5 high values (darker) and 20 low values (lighter). Each of these arrays corresponds to one of the topics distributions in our toy problem, that is, each one places the majority of its probability mass on a different set of 5 symbols.

The $\boldsymbol{\gamma}_1 \ldots \boldsymbol{\gamma}_K \stackrel{\circ}{=} \boldsymbol{\gamma}$ parameters in the HMTM are a set of K-dimensional Dirichlet parameters. Each γ_k is an array of K non-negative real numbers. A common reparameterization of Dirichlet parameters, such as γ_k is as $s\mathbf{m}_k$, where s is the sum of γ_k and $\mathbf{m}_k$ is the γ_k divided by s. For each set of K Dirichlet parameters, we used an $s = 3.0$. The K arrays $\mathbf{m}_1 \ldots \mathbf{m}_K$ are depicted below on the left.



From these Dirichlet parameters, we may generate arbitrarily many sets of state-transition parameters for a 10-state HMM. As an example, we show two such sets on the right. As is evident, these parameters retain characteristics of the patterns found in the original Dirichlet parameters. We can see that, on average, the state transition dynamics leads a given state to map to either the state before it, or the state after it. For example, we can see that, on average, state 2 maps to state 1 or state 3, state 3 maps to state 2 or state 4, and so on. While this average dynamical behavior is simple, it is not trivial and does not lead to fixed point or periodic trajectories. Note also that the small differences in the transition dynamics can lead to quite distinct dynamical behaviors in their respective HMMs.

On the basis of these $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, and using an array of K ones as the parameters $\boldsymbol{\alpha}$, we generated $J = 50$ training ''texts'' as follows. For each text j, we drew a set of initial conditions and transition parameters for a HMM from $\boldsymbol{\alpha}$ and γ, respectively. We then iterated the HMM for $n_j = 100$ steps, to obtain a state trajectory $x_{j1}\ldots x_{jn_j}$. On the basis of the value of each x_{ji}, we chose the appropriate topic (i.e., if $x_{ji} = k$ we chose $\boldsymbol{\phi}_k$) and drew an observation w_{ji} from it. The training thus comprised a set of J symbol sequences, with each symbol taking a value from the set $\{1\ldots25\}$.

Using this as training data, we trained another HMTM whose parameters were unknown. As described earlier, the Gibbs sampler draws samples from the posterior distribution over $\mathbf{x}_{1:J}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and γ. From these samples, we may also draw samples from the posterior over the topics. In Fig. 3, we graphically present some results of these simulations. Show in this figure are averages of samples drawn from the posterior over $\boldsymbol{\phi}$ and $\mathbf{m}_1\ldots\mathbf{m}_k$, that is, the location parameters of $\gamma_1\ldots\gamma_k$. On the top row of Fig. 3, we show averages of samples drawn after 20 iterations of the sampler. On the lower row, we show averages of samples drawn after 100 iterations. In both cases, these averages are over 10 samples taken two iterations apart in time. To the left in each case are the inferred topics. To the right are the inferred locations of the Dirichlet parameters. These inferred parameters can be compared to the true parameters on the previous page. By doing so, it is clear that even after 20 iterations of the sampler, patterns in the topic distributions have been discovered. After 100 iterations, the inferred parameters are almost identical to the originals.

Although not shown, the Gibbs sampler also successfully draws samples from the posterior over $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ [2] and the scale parameter s for γ. In addition, we may use draws from the posterior to estimate, using the harmonic mean method, the marginal likelihood of the model under a range of different numbers of topic distributions. Although, the harmonic mean method is not highly recommended, we have found that in practice it consistently leads to an estimate of the correct number of topics.

4. Learning topics from text

In this final section, we present some topics learned by a HMTM trained on a corpus of natural language. The corpus used was a subsample from the British National Corpus (BNC).[3] The BNC is annotated with the structural properties of a text such as sectioning and subsectioning information. The latter type of annotation facilitates the processing of the

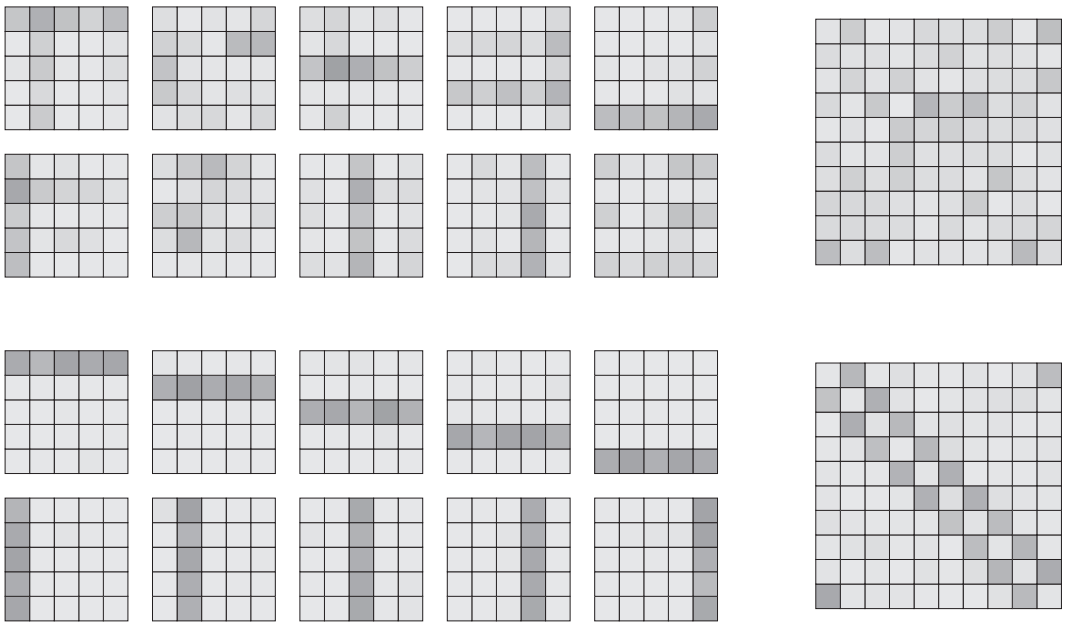


Fig. 3. Averages of samples from the posterior distribution over the topic distributions (left) and locations of the γ parameters (right). Shown on the top row are averages, over 10 draws, drawn after 20 iterations. Shown on the lower row are averages, again over 10 draws, taken after 100 iterations. Compare with the true parameters shown on the previous page.

corpus. To extract texts from the BNC we extracted contiguous blocks that were labeled as high-level sections, roughly corresponding to individual articles, letters, or chapters. These sections varied in size from tens to thousands of words, and from these we chose only those texts that were approximately 150–250 words in length. This length is typical of, for example, a short newspaper article. Following these criteria, we then sampled 2,500 individual texts to be used for training. Of all the word types that occurred within this subset of texts, we excluded words that occurred less than five times overall and replaced their occurrence with a marker symbol. This restriction resulted in a total of 5,182 unique words.

We trained a HMTM with $K = 120$ topics using this corpus. After a burn-in period of 1,000 iterations, we drew 50 samples from the posterior over the latent trajectories and $\boldsymbol{\beta}$, with each sample being 10 iterations apart. We used these to draw samples from the posterior over the topics, which are then averaged, as described earlier. In the upper part of Table 2, we present seven averaged topics from the HMTM simulation. For the purposes of comparison, in the lower part Table 2 we present seven averaged topics taken from a standard Topics model. This standard Topics model was trained on a larger corpus and is described in detail in Andrews et al. (2009). The topics in the standard model were chosen by finding the topics that are the closest matching to the HMTM topics we chose.

The side-by-side comparison provides an appreciation for how the topics in a HMTM differ from the standard model. In the HMTM, the topics are more refined in the semantics, referring to more specific categories of things or events. For example, we see that the first

Table 2
Examples of topics learned by a HMTM (top row) that was trained on a set of documents taken from the BNC

HMTM Topics						
beer	sheep	sugar	aircraft	film	say	ship
guinness	cattle	fruit	plane	movie	know	boat
alcohol	meat	butter	jet	series	talk	ferry
ale	livestock	bread	airline	tv	think	vessel
whisky	dairy	chocolate	squadron	story	feel	ships
spirits	beef	milk	helicopter	television	understand	navy
wine	pigs	cream	fighter	soap	believe	shipping
cider	animal	water	hercules	movies	speak	lifeboat
pint	cow	lemon	airbus	drama	ask	fleet
lager	pig	egg	falcon	episode	explain	coastguard
Standard Topics						
pub	farm	fruit	air	film	want	boat
drink	agriculture	add	aircraft	star	think	island
guinness	food	fresh	flight	hollywood	like	sea
beer	farming	butter	plane	movie	people	ship
drinking	sheep	cooking	airport	screen	moment	crew
wine	agricultural	minutes	flying	stars	happen	ferry
bar	cattle	hot	pilot	director	wanted	sailing
alcohol	ministry	food	fly	actress	worried	yacht
brewery	crop	bread	jet	actor	believe	shipping
whisky	pigs	chicken	airline	role	exactly	board

Note. On the lower row, we show topics from a standard Topics model also trained on a set of documents from the BNC.

topic to the left in the HMTM refers to alcoholic beverages, specifically those associated with a (British) pub. By contrast, the corresponding topic from the standard model is less specifically about beverages and refers more generally to things of, or relating to, pubs. In the next example, we see that the topic from the HMTM refers to farm animals. By contrast, the corresponding topic from the standard model is less specifically about farm animals and more related to agriculture in general. In all of the examples shown, a similar pattern of results holds.

From this demonstration, we can see that more refined semantic representations can be learned when sequential information is taken into account. Why we see a benefit from sequential information can be understood by way of the following simple example. Words like *horse*, *cow*, *mule* are likely to occur as subjects of verbs like *eat*, *chew*, while words like *grass*, *hay*, *grain* are likely to occur as their objects. A model that learns topics by taking into account this sequential information may learn that words like *horse*, *cow*, and *mule* etc., form a coherent topic. Likewise, such a model may infer other topics based on words like *eat*, *chew*, etc., or words like *grass*, *hay*, *grain*, etc. By contrast, the standard Topics model, based on the assumptions that the sequential information in a text is irrelevant, is likely to conflate these separate topics into one single topic referring to, for example, *farms* or *farming*.

5. Discussion

The general problem that has motivated the work presented here is the problem of how word meanings are represented and learned. As mentioned, a promising recent approach to this general problem has been the study of how the meanings of words can be learned from their statistical distribution across the language. Across disciplines such as cognitive science, computational linguistics, and natural language processing, the various computational approaches to modeling the role of distributional statistics in the learning of word meanings can be seen to fall into two broad groups, each characterized largely by the granularity of statistical information that they consider. On the one hand, there are those models that focus primarily on broad or coarse-grained statistical structures. These models, which include the seminal work of, for example, Schütze (1992); Lund et al. (1995); Landauer and Dumais (1997), have been widely adopted and have many appealing characteristics. In most cases, they are large-scale, broad coverage, unsupervised learning models that are applied to plain and untagged corpora. In addition, particularly in recent models such as Griffiths et al. (2007), they are often also fully specified probabilistic generative language models. However, one of the widely shared practices of models like this is to adopt a bag-of-words assumption and treat the linguistic contexts in which a word occurs as unordered sets of words. The obvious limitation of this approach is that all fine-grained sequential and syntactic statistical information is disregarded.

In contrast to coarse-grained distributional models, a second group of models have focused specifically on the role of more fine-grained sequential and syntactic structures (e.g., Padó & Lapata, 2007; Pereira, Tishby, & Lee, 1993). The obvious appeal of these models is that by focusing on the sequential order, argument structure, and general syntactic relationships, within sentences, they capture syntactic-semantic correlations in language that can provide vital information about the possible meanings of words (see, e.g., Alishahi & Stevenson, 2008; Gillette, Gleitman, Gleitman, & Lederer, 1999). However, in practice, these fine-grained distributional models have often focused on extracting semantic information for constrained or specific problems (e.g., automatic thesaurus generation; Curran & Moens, 2002a,b; Lin, 1998) or taxonomy generation (Widdows, 2003) rather than developing large-scale probabilistic language models. This has limited their suitability as cognitive models. In addition, they have been heavily reliant on supervised learning methods using part-of-speech tagged or parsed corpora.

In this paper, we have described the HMTM, a model that attempts to integrate the coarse- and fine-grained approaches to modeling distributional structure. This model is intended to preserve the strengths of both of these approaches, while avoiding their principal limitations. It is scalable, unsupervised, and trained with untagged corpora. However, it goes beyond the dominant bag-of-words paradigm by incorporating sequential and syntactic structures. In this respect, it is similar to the BEAGLE model described in Jones and Mewhort (2007). However, unlike BEAGLE, the HMTM, being a direct generalization of the LDA model, is a fully specified probabilistic generative language model. The HMTM also extends beyond a model described in Griffiths et al. (2007);

Griffiths, Steyvers, Blei, and Tenenbaum (2005) that couples a HMM with a standard Topics model. In particular, the HMTM is designed to learn semantic representations by directly availing of the sequential information in text. By contrast, the model described in Griffiths et al. (2005, 2007) learns semantic representations using a standard Topics model (i.e., latent variables are drawn by independently sampling from a fixed distribution), while the HMM is used to learn syntactic categories. More generally, the HMTM is an example of a hierarchical language model. Within cognitive science, computational linguistics, and NLP research, probabilistic language models such as HMMs and PCFGs have been extensively used. However, the hierarchically coupled versions of these models have yet to be introduced. The HMTM is a hierarchically coupled HMM and can be in principle extended to the case of a hierarchically coupled PCFG. We believe that these models, and their training by Bayesian methods on large-scale data-sets, may be an important development in the complexity of the language models used in these fields.

Notes

1. The $\boldsymbol{\beta}$ parameters can be seen as a prior over $\boldsymbol{\phi}$, but a prior that will be inferred from the data rather than simply assumed.
2. For the case of $\boldsymbol{\beta}$, we used a symmetric Dirichlet distribution.
3. The BNC is a 100-million-word corpus of contemporary written and spoken English in the British Isles. According to its publishers, the texts that make up the written component include ''extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays.''

Acknowledgments

We would like to thank Mark Steyvers, Roger Levy, Nick Chater, and Yee Whye Teh for their valuable comments and discussion. This research was supported by European Union (FP6-2004-NEST-PATH) grant 028714.

References

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, *32*, 789–834.

Andrews, M., Vigliocco, G., & Vinson, D. P. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463–498.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boyd-Graber, J., & Blei, D. (2009). Syntactic topic models. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems* (pp. 185–192). Cambridge, MA: MIT Press.

Curran, J. R., & Moens, M. (2002a). Improvements in automatic thesaurus extraction. In *Proceedings of the workshop on unsupervised lexical acquisition* (pp. 59–66). Philadelphia, PA: Association for Computational Linguistics.

Curran, J. R., & Moens, M. (2002b). Scaling context space. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 231–238). Morristown, NJ: Associations for Computational Linguistics.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in linguistic analysis (Special volume of the philological society, Oxford)* (pp. 1–32). Oxford, England: Blackwell.

Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, *41*(2), 337–348.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176.

Griffiths, T., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In W. D. Gray & C. Schunn (Eds.), *Proceedings of the 24th annual conference of the cognitive science society* (pp. 381–386). Mahwah, NJ: Erlbaum.

Griffiths, T., & Steyvers, M. (2003). Prediction and semantic association. In S. T. S. Becker & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 11–18). Cambridge, MA: MIT Press.

Griffiths, T., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In L. K. Saul, Y. Weiss, & L. Botlou (Eds.), *Advances in neural information processing* (Vol. 17, pp. 537–544). Cambridge, MA: MIT Press.

Griffiths, T., Steyvers, M., & Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Harris, Z. (1954). Distributional structure. *Word*, *10*(2–3), 775–793.

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

Landauer, T., & Dumais, S. (1997). A solutions to Plato's problem: The Latent Semantic Analyis theory of acquistion, induction and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the joint annual meeting of the association for computational linguistics and international conference on computational linguistics* (pp. 768–774).

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In J. D. Moore & J. F., Lehman (Eds.), *Proceedings of the seventeeth annual conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161–199.

Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In *31st annual meeting of the ACL* (pp. 183–190). Association for computational Linguistics .

Schütze, H. (1992). Dimensions of meaning. In *Proceedings of supercomputing* (pp. 787–796). IEEE Computer Society Press.

Wallach, H. (2006). Topic modeling: Beyond bag-of-words. In W. W. Cohen & A. Moore (Eds.), *Proceedings of the 23rd international conference on machine learning* (Vol. 148, pp. 977–984). ACM.

Widdows, D. (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In M. Hearst & M. Ostendorf (Eds.), *Naacl '03: Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology* (pp. 197–204). Morristown, NJ: Association for Computational Linguistics.

Appendix

The Gibbs sampler for the HMTM model draws samples from $P(\mathbf{x}_{1:J},\boldsymbol{\alpha},\boldsymbol{\beta},\gamma \mid \mathbf{w}_{1:J})$. It does so iteratively sampling from a given latent variable x_{ji}, assuming values from all other latent variables, and for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, γ. The conditional distribution over x_{ji} is given by

$$\begin{aligned} &P(x_{ji}|\mathbf{x}_{-[ji]}, \mathbf{w}_{1:J}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \\ &\propto P(w_{ji}|x_{ji}, \mathbf{x}_{-[ji]}, \mathbf{w}_{-[ji]}, \boldsymbol{\beta}) P(x_{ji}|\mathbf{x}_{-[ji]}, \boldsymbol{\alpha}, \boldsymbol{\gamma}), \end{aligned} \tag{A.1}$$

where we denote the set of latent variables excluding x_{ji} by $\mathbf{x}_{-[ji]}$, and denote the set of observables excluding w_{ji} by $\mathbf{w}_{-[ji]}$. Superficially, this conditional distribution appears identical to the conditional distributions in the Gibbs sampler for the standard Topics model, as described in Griffiths and Steyvers (2002, 2003); Griffiths et al. (2007). However, due to the non-independence in the latent trajectory that results from the Markov dynamics, the term $P(x_{ji} \mid \mathbf{x}_{-[ji]},\boldsymbol{\alpha},\gamma)$ must be calculated as a ratio of Polya distributions, that is,

$$P(x_{ji}|\mathbf{x}_{-[ji]}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \frac{P(\mathbf{x}_{1:J}|\boldsymbol{\alpha}, \boldsymbol{\gamma})}{P(\mathbf{x}_{-[ji]}|\boldsymbol{\alpha}, \boldsymbol{\gamma})}. \tag{A.2}$$

The Dirichlet parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, γ may also be sampled by Gibbs sampling. For example, each γ_k is reparameterized as $s\mathbf{m}_k$ (as described in the main text). Assuming known values for all variables, the conditional posterior distribution of s is log-concave and can be sampled using adaptive rejection sampling (ARS), see Gilks and Wild (1992). Likewise, assuming known values for all other variables, the conditional posterior over the share of the probability mass between any two elements of $\mathbf{m}_k$ is also log-concave and can be sampled by ARS. As such, the Gibbs sequentially samples from each latent variable x_{ji}, each scale parameter of the Dirichlet parameters given by $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, γ, and also from the share of probability mass between every pair of elements of each location parameters of the Dirichlet parameters.