The Role of Attributional and Distributional Information in Semantic Representation

Mark Andrews (mark@gatsby.ucl.ac.uk)
Gatsby Computational Neuroscience Unit, University College London
London, WC1N 3AR

Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)
David Vinson (d.vinson@ucl.ac.uk)
Department of Psychology, University College London
London, WC1E 6BT

Abstract

In recent studies of semantic representation, two distinct sources of information from which we can learn word meanings have been described. We refer to these as *attributional* and *distributional* information sources. Attributional information describes the attributes or features associated with referents of words, and is acquired from our interactions with the world. Distributional information describes the distribution of words across different linguistic contexts, and is acquired through our use of language. While previous work has concentrated on the role of one source, to the exclusion of the other, in this paper we study the role of both sources in combination. We describe a general framework based on probabilistic generative models for modelling both sources of information, and how they can be integrated to learn semantic representation. We provide examples comparing the learned structures of each of three models: attributional information alone, distributional information alone, and both sources combined.

Keywords: Probabilistic Models; Semantic Representation; Information Integration.

Introduction

We can distinguish between two distinct sources that provide us with information about the meanings of words. The first is what we will call *attributional* information. This describes physical, emotional and conceptual (or otherwise nonlinguistic) attributes associated with the referents of words. These are features or attributes associated with words that are based upon or built up from our interactions in the environment, and our knowledge of the objects and relationships in the world. A second source of information about word meanings is what we will call *distributional* information. This describes the distribution of words across different linguistic or textual contexts. This type of information has been memorably summarized by the phrase "you shall know a word by the company it keeps", (Firth, 1957). In other words, the type of linguistic contexts in which a word occurs can provide clues as to what that word might mean.

While these two sources of information are not uncorrelated, neither are they identical, and it is plausible to assume that they are both utilized when learning the meaning of words. The two sources are correlated because words that refer to similar things or events in the world are likely to appear in similar linguistic contexts. The reverse case could also be argued. Knowing that two words appear in similar texts might imply that these words refer to similar things in the world. It is, however, reasonable to assume that these sources are distinct, or that one source is not entirely explained by or dependent upon the other. For example it can be argued that attributional information is more important in learning words referring to concrete entities and actions, for which the properties of the thing or event in the world can be experienced via the senses, whereas distributional information may be more important in learning abstract words, those that we learn primarily via linguistic context. Thus both types of information can be exploited in order to learn word meanings. In previous studies, the contributions of either source alone, independent of the other, have been studied. In particular, within cognitive psychology, researchers have primarily focused on the development of models of meaning representation based on attributional information (Collins & Quillian, 1969; Hinton & Shallice, 1991; McRae, Sa, & Seidenberg, 1997; McClelland & Rogers, 2003; Minsky, 1975; Smith, Shoben, & Rips, 1974; Vigliocco, Vinson, Lewis, & Garrett, 2004), whereas recently within computational linguistics, machine learning and related areas in cognitive science, researchers have primarily focused on distributional information alone (Burgess & Lund, 1997; Dagan, Lee, & Pereira, 1997; Griffiths & Steyvers, 2003; Hofmann, 1999b; Landauer & Dumais, 1997; Schütze, 1992). In this paper, we describe a model that uses both sources in combination as the basis of semantic representation.

Generative Models of Semantic Representations

Probabilistic generative models provide a general means by which to model semantic representation. This approach has already been pursued both within machine learning (Blei & Ng, 2003; Hofmann, 1999b, 1999a; Teh, Jordan, Beal, & Blei, 2004) and within cognitive science (Griffiths & Steyvers, 2003, 2002). Generative models describe the data of interest in terms of explicit probabilistic relations and variables. The nature of these relationships and variables are inferred or learned from data. This general class of models can be used to model the role of both attributional and distributional information. Indeed, many of the previous models of semantic representation, mentioned above, can be described as implicitly falling within this general class. In this paper, we will use this approach to model both the independent and joint role of distributional and attributional information in semantic representations.

Considering the role attributional information alone plays in semantic representation, a suitable generative model would describe how concrete terms (or more properly, their referents) are generated. In this case, a simple generative model might assume that concrete terms are defined in terms of a distribution over latent attribute classes, and that these attribute classes are themselves probability distributions over binary properties or features. Concrete terms are distributions over these features, obtained by repeatedly sampling from the distribution of latent attribute classes and then sampling from the distributions over binary properties associated with these classes.

Considering the role distributional information alone plays in semantic representation, an appropriate generative model would describe how a *text* is generated. Henceforth, we will use the term *text* to refer to any linguistic or textual utterance. While this could be an entire article, book, or transcribed conversation, we will usually use the term to refer to paragraphs, sentences, or strings of consecutive words in written documents or transcribed speech. A simple generative model of texts might assume that texts are multinomial distributions over latent semantic classes or topics, and that these latent topics are themselves multinomial distributions over words. A text is generated by repeatedly sampling from the distribution over latent topics, and then sampling from their corresponding distributions over words. In such a model, the semantic representation of a given word is defined in terms of its posterior probability distribution over the latent classes. In other words, the semantic representation of a given word is the probability distribution over the latent topics that can be inferred whenever the word is generated.

Combining the two sources of information is straightforward. The data we observe consists of sets of words, and associated with each word is a distribution over binary non-linguistic attributes. As will be clarified, we can assume that a distribution over latent variables accounts for both the distribution of words in a text and the distribution of binary features associated with a given word. The semantic representation of a word is defined in terms of its posterior distribution over these latent topics. These semantic representations will be constrained to account for both the distributions of words in texts, and the distributions of binary attributes associated with given words.

The intuitive rationale behind the model is as follows: On its own, the statistical structure and patterns in a language can provide information about word meanings. A subset of the words in the language are associated with physical objects or events in the world, and this information can be integrated with statistical patterns in language. As a contrived example, knowledge that the word *cat* refers to those creatures with claws and whiskers that meow, etc. can be integrated with implicitly acquired statistical knowledge that *cat* co-occurs with the words *dog* and *pet*, etc. This two sources of information could be combined to provide a richer understanding of the semantics of the word *cat* than could be learned by either source alone.

The Model: Formal Specifications

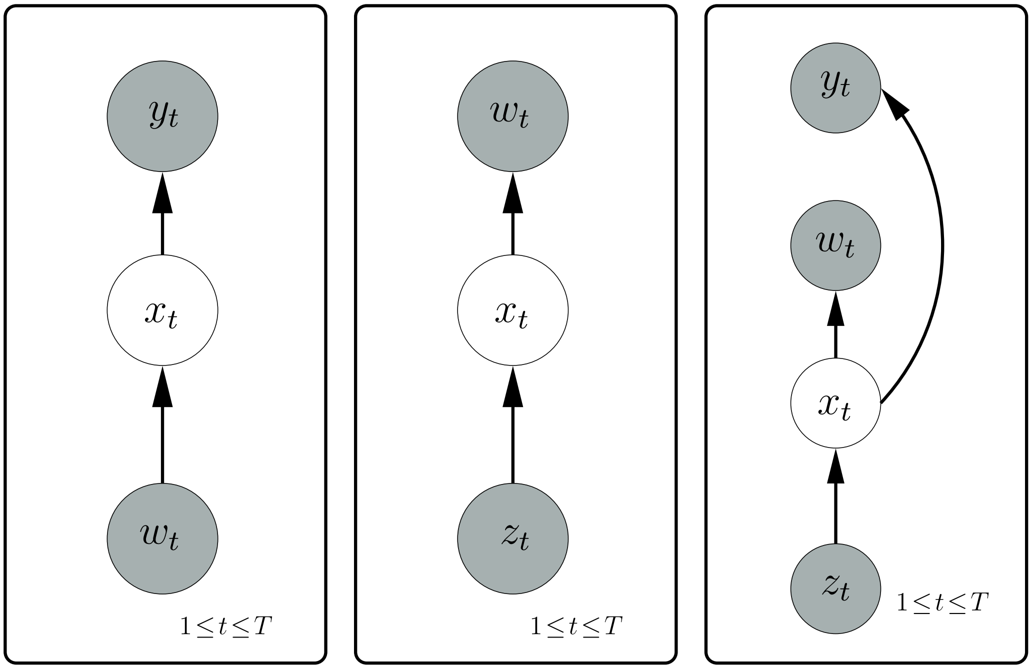
We can make the preceding description more formal as follows. The observable data that we are modelling consist of both texts and the attributes associated with words. As mentioned, texts take the form of paragraphs, sentences and strings of consecutive words in a natural language corpus. If there are J texts in a corpus, we can label them as $\{z_1, z_2 \ldots z_j \ldots z_J\}$. The texts are, in general, of variable length. For example, text z_j is of length T_j and consists of the sequence of words $\{w^j_1, w^j_2 \ldots w^j_{t_j} \ldots w^j_{T_j}\}$. For each possible j and t, $w^j_{t_j}$ will be a word in the vocabulary of word-types $\mathcal{V} = \{v_1, v_2 \ldots v_k \ldots v_K\}$. What defines a word-type for our purposes is described below, but in general, word-types are a set of common dictionary words in English.

A subset of the words in any human language will *concrete words*, or words that refer to objects, events, actions, etc., in the world. Assume that of the K dictionary words in $\mathcal{V}$ we have obtained a set $\mathcal{V}_f$ of N concrete words, where $N \leq K$. For each word in $\mathcal{V}_f$ we have a probability distribution over a set of L binary features. We can represent this binary feature vector by $y = \{y_1, y_2 \ldots y_l \ldots y_L\}$, where y_l is the lth binary feature of y. Each feature is a property or attribute that could be associated with the physical referent of the concrete words in $\mathcal{V}_f$.

To model semantic memory according to the description provided above, we provide generative models for the case where attributional data alone is used, where the distributional data alone is used, and where both data sources are used. The graphical models (or Bayesian Networks) for these three generative models are shown in Figure 1. A graphical model describes the conditional independence structure of the variables in a model. Taken together, we have the observable variables w_t, y_t and z_t that represent, respectively, the words, features and text occurring at a time t. In addition, we introduce the latent-variable $x_t \in \{\xi_1, \xi_2 \ldots \xi_m \ldots \xi_M\}$. As a latent, or hidden, variable the value of x_t is unobserved. We see that in the attributional model, the binary attribute vector y_t is conditioned upon x_t, while x_t is conditioned upon the word label w_t. In the distributional model, we have the observable words conditioned upon the latent-variable x_t, which is then conditioned upon the text z_t. In the combined model, both the words w_t and binary attribute vector y_t are conditioned upon the latent-variable x_t, with x_t conditioned upon the text z_t. The parametric forms of these conditional distributions are as follows. (The parameters in the models are referred to generically as θ, until further specified).

Attributional Model

In the attributional model, the probability of the observed binary attributes y_t, conditioned upon w_t can be



(a) Attributional Model (b) Distributional Model (c) Combined Model

Figure 1: The generative models that utilize a) attributional, b) distributional and c) combined information sources.

factored as

$$\mathrm{P}\big(y_t \big| w_t, \theta\big) = \sum_{\{x_t\}} \mathrm{P}\big(y_t \big| x_t, \theta\big)\, \mathrm{P}\big(x_t \big| w_t, \theta\big)\,, \qquad (1)$$

where the latent variable x_t is integrated over. The probability that the l^{th} binary variable of y_t, i.e y_l^t, takes the value 1 given that $x_t = \xi_m$, is denoted by p_{ml}. The probability that x_t takes the value ξ_m given the value of w_t, is denoted by $\pi_m^{[w_t]}$. We can re-write (1) as

$$\mathrm{P}\big(y_t \big| w_t, \theta\big) = \sum_{m=1}^{M} \pi_m^{[w_t]} \prod_{l=1}^{L} p_{ml}^{(1-y_l^t)} (1 - p_{ml})^{(1-y_l^t)}. \qquad (2)$$

As such, the attributional model is a mixture of M multivariate Bernoulli distributions.

Distributional Model

In the distributional model, the probability of the observed word w_t, conditional upon observing text z_t can be factored as

$$\mathrm{P}\big(w_t \big| z_t, \theta\big) = \sum_{\{x_t\}} \mathrm{P}\big(w_t \big| x_t, \theta\big)\, \mathrm{P}\big(x_t \big| z_t, \theta\big)\,. \qquad (3)$$

The probability of observing that $w_t = v_k$ given that $x_t = \xi_m$ is denoted by q_{mk}. The probability that $x_t = \xi_m$ upon observing that $z_t = j$, is denoted by $\pi_m^{[z_t]}$. We can re-write (3) as

$$\mathrm{P}\big(w_t \big| z_t, \theta\big) = \sum_{m}^{M} \pi_m^{[z_t]} \prod_{k=1}^{K} q_{mk}^{\mathbb{I}(w_t = v_k)}, \qquad (4)$$

where $\mathbb{I}(a)$ is an indicator function, taking the value 1 if its argument a is true, and 0 otherwise. The distributional model is, as such, a mixture of multinomial distributions.

Combined Model

In the combined model, as shown in Figure 1, both the word label w_t and attribute vector y_t are conditioned upon the latent variable x_t, which is conditioned upon the text z_t. Integrating over the values of x_t, the conditional likelihood of the observables is

$$\begin{aligned} &\mathrm{P}\big(y_t, w_t \big| z_t, \theta\big) \\ &= \sum_{\{x_t\}} \mathrm{P}\big(y_t \big| x_t, \theta\big)\, \mathrm{P}\big(w_t \big| x_t, \theta\big)\, \mathrm{P}\big(x_t \big| z_t, \theta\big) \end{aligned} \qquad (5)$$

and substituting in parameters p_{ml}, q_{mk} and $\pi_m^{[z_t]}$ as in (2) and (4)

$$= \sum_{m}^{M} \pi_m^{[z_t]} \prod_{k=1}^{K} q_{mk}^{\mathbb{I}(w_t = v_k)} \prod_{l=1}^{L} p_{ml}^{(1-y_l^t)} (1 - p_{ml})^{(1-y_l^t)}. \qquad (6)$$

As such, the combined model is a mixture of both multivariate Bernoulli distributions and multinomial distributions.

Model Learning

Given a set of training data $\mathcal{D}$ consisting of both texts and attributes associated with concrete words, for each model above we would ideally wish to estimate $\mathrm{P}\big(\theta\big|\mathcal{D}\big)$, or the posterior probability of the parameters given the data. In any analyses of the models, we would then integrate over this entire distribution. In related work, we are using Markov-Chain Monte Carlo (MCMC) simulations to compute these posteriors, however these studies will not be described here. For present purposes, we will approximate $\mathrm{P}\big(\theta\big|\mathcal{D}\big)$ by its modal point θ_{mp}, which assuming a prior distribution over the parameters, is given by its maximum likelihood estimate θ_{mle}.

The standard procedure for maximum-likelihood (or maximum posteriori) estimation in latent-variable models is Expectation-Maximization (EM). This consists of iteratively computing a lower-bound on the likelihood of the data $\mathrm{P}\big(\mathcal{D}\big|\theta\big)$, and maximizing this bound with respect to the parameters. This leads to a set of parameter update rules that can be guaranteed to monotonically increase the likelihood and converge to an (at least local) maximum. For example, in the case of the combined model above, the update rules for p_{ml}, q_{mk} and $\pi_m^{[z_t]}$ are

$$p_{ml} \propto \sum_{j=1}^{J} \sum_{t_j=1}^{T_j} \mathrm{P}\big(x_t = \xi_m, \big| w_t, y_t, z_t = j, \theta\big)\, y_l^t, \tag{7}$$

$$q_{mk} \propto \sum_{j=1}^{J} \sum_{t_j=1}^{T_j} \mathrm{P}\big(x_t = \xi_m, \big| w_t, y_t, z_t = j, \theta\big)\, \mathbb{I}(w_t = v_k), \tag{8}$$

$$\pi_m^{[z_t]} \propto \sum_{t_j=1}^{T_j} \mathrm{P}\big(x_t = \xi_m, \big| w_t, y_t, z_t = j, \theta\big)\,. \tag{9}$$

The update rules for the attributional model and distributional model are special cases of the above rules, with the appropriate changes having been made.

Simulations

The text corpora used consisted of fiction and non-fiction books publicly available at the Oxford Text Archive ($\approx 6.5 \cdot 10^6$ words) and Project Gutenberg ($\approx 11.6 \cdot 10^6$ words), Reuters Newswire texts ($\approx 2.5 \cdot 10^6$ words), and a set of Usenet articles ($\approx 5.25 \cdot 10^6$ words). We folded British into American spellings (e.g. *centre* to *center*, *favour* to *favor*, *realised* to *realized*, etc.), folded affix-variations of words into their word-stems (e.g. *swims*, *swam*, *swum*, *swimming* changed to *swim*, etc.), eradicated non-words (using a standard American-English dictionary), and eradicated stop-words (using a standard list of ≈ 550 stop-words). This reduced the corpora to a total size of $\approx 7.7 \cdot 10^6$ words, with $16,956$ word-types. By further eradicating all word-types that appear with a frequency of greater than 10^4 or less than 10^2, we can reduce the total size to $\approx 6.1 \cdot 10^6$ words, and $7,393$ word-types. This corpus was divided into a set of $51,160$ texts, each of which were ≈ 150 words long.

In a previous study, Vigliocco et al. (2004) compiled frequencies of a set 1029 binary attributes associated with a set of 456 common words. The 456 words consisted of 230 nouns, of which 169 referred to objects and 71 to actions, and 216 verbs, all of which referred to actions. The attribute-types and their frequencies were collected from speaker-generated lists of attributes associated with the 456 words. This was done in a manner similar to that used in McRae et al. (1997). Certain word types referred to distinct verb and noun senses, e.g. *(the) hammer* and *(to) hammer*. For present purposes, these word-types were regarded as identical and the vectors associated with the words were collapsed. Out of this set of words, a subset of exactly 300 also occurred in our reduced (i.e. $7,393$ word-types) text corpora.

In summary, the text corpora consist of $51,160$ texts, each described by a frequency distribution over $7,393$ word-types. Of these $7,393$ word-types, a subset of 300 are also described by a frequency distribution over 1029 attribute-types. The word-attribute set alone, the text set alone and the word-attribute and text-set together were used to train the distributional model, attributional model and combined models, respectively. In all cases, this was done using EM to find the maximum likelihood estimates of the parameters given the observed data-sets. In the simulations described here, the attributional model used a latent-variable of dimensionality 150 with both the distributional model and the combined model used latent-variables of dimensionality 300.

Analysis of Trained Models

As described above, the latent variables in each model can be seen as distributions over words, attributes or both. In Figure 2, we illustrate some of these latent variables. For each case, we draw samples of the words and/or attributes with which they are associated. The distributional model learns latent variables that correspond to multinomial distributions over words. These distributions can be intuitively viewed as *topics* that have been learned by the model. In the three example latent variables shown for the distributional model (i.e. the three leftmost columns of Figure 2), we see topics that could be labelled *government*, *business* and *religion*. In the attributional model, latent variables correspond to distributions over binary attributes and can intuitively be viewed as attribute classes. Examples of attribute-classes, and the samples over attributes to which they correspond, are shown in the middle three columns of Figure 2). These classes could be labelled *human-body*, *fruit-vegetable* and *clothing*. In the case of the combined model, latent variables correspond to distributions over both words and attributes. In this sense, they are merges or combinations of both attributional and distributional classes. We provide some examples in the rightmost three columns of Figure 2). The classes learned could be labelled *transport*, *medical* and *war*, each defined both by clusters of words and clusters of attributes. The words are given in uppercase, while the attributes are in lowercase.

In each model, we can measure how much any given

NATION	MONEY	ALLAH	human	fruit	leg	CAR	PATIENT	WAR
AUTHORITY	HUNDRED	BIBLE	face	green	clothing	HORSE	MEDICAL	GUN
PRINCIPLE	COURT	BELIEF	hair	grow	wear	RIDE	DOCTOR	KILL
CENTURY	LAND	APOSTLE	eye	red	body	DRIVE	HEALTH	ATTACK
UNITE	PAY	CHURCH	shoulder	round	protect	DRIVER	MEDICINE	KNIGHT
GOVERNMENT	THOUSAND	DISBELIEVE	leg	sweet	cover	transport	nose	kill
SOCIETY	TAX	CHRIST	hand	juice	body	vehicle	body	weapon
RELIGION	CITY	JESUS	body	tree	warm	4-legs	human	anger
CONSTITUTION	SCIENCE	MARRY	foot	eat	long	wheel	eye	fear
POLITICAL	OFFICE	SPIRIT	mouth	food	humans	car	head	yell

Figure 2: Examples of latent classes learned by (from left to right) the distributional, attributional and combined models. For each example latent class, we have drawn samples of the words and/or the attributes they correspond to. Capitals refer to words, while lower case refer to attributes.

word predicts any other. The extent to which word v_j predicts v_i is given by

$$\begin{aligned}&\mathrm{P}\big(w = v_i\big|w = v_j, \theta\big)\\ &\quad= \sum_{\{x\}} \mathrm{P}\big(w = v_i\big|x, \theta\big)\,\mathrm{P}\big(x\big|w = v_j, \theta\big)\,.\end{aligned} \tag{10}$$

In both the distributional model and the combined model, $\mathrm{P}\big(w = v_i\big|x, \theta\big)$ is the likelihood term of the model, while $\mathrm{P}\big(x\big|w = v_j, \theta\big)$ is given by Bayes rule,

$$\mathrm{P}\big(x\big|w = v_j, \theta\big) \propto \sum_{\{z\}} \mathrm{P}\big(w = v_j\big|x\big)\,\mathrm{P}(x, z)\,. \tag{11}$$

On the other hand, for the attributional model $\mathrm{P}\big(x\big|w = v_j, \theta\big)$ is directly available, while $\mathrm{P}\big(w = v_i\big|x, \theta\big)$ is obtained by Bayes rules, i.e.

$$\mathrm{P}\big(w = v_i\big|x, \theta\big) \propto \mathrm{P}\big(x\big|w = v_i, \theta\big)\,\mathrm{P}\big(w = v_i\big|\theta\big) \tag{12}$$

How predictive any word is, given an observation of another word, can be taken as a measure of semantic relatedness. This measure may, perhaps, be more theoretically motivated than other measures that are based upon a distance metric. This argument is already presented in detail in Griffiths and Steyvers (2003). What a prediction of a word v_i by a word v_j means can be interpreted as *given that we are observing v_j in this text, how probable is word v_i*. Below, we show words predicted by some example words in each of the three models. For comparison purposes between the models, here we provide only prediction of words that were in both the text-based and attribute-based data-sets.

Dog

Distributional: growl bark chase lick whine cat tail paw wolf snap

Attributional: cat rabbit goat tail pig fox sheep horse bear fur

Combined: cat growl tail bark paw whine chase sheep lick wolf

Gun

Distributional: threat stab knife bomb kick kill argue snap knock murder

Attributional: murder sword dagger bomb threat knife shield threaten stab scream

Combined: threat knife murder stab bomb threaten kick snap knock argue

Ride

Distributional: motorcycle bicycle horse chase pant slide thumb truck clatter ankle

Attributional: carry drive travel train move pull approach walk push truck

Combined: motorcycle bicycle horse travel pull slide truck chase carry push

Discussion

The attributes associated with a word, and its distribution across texts both provide information about its the meaning. Either source alone can provide information not provided by the other: Attributes provide information about the physical (or nonlinguistic) relationships in the world between the referents of the word, while distributional information provide a rich patterns of linguistic (or non-physical) contexts where the word is found. While attributional information is undoubtedly used to learn concrete words, distributional information provides valuable information about the behavior of abstract words. While in previous studies, the role in semantic representations of each source, independent of the other, has been investigated, in this paper we describe how both sources can be used in combination.

In using both sources of information, patterns in one source can interact with those of another. As a trivial example, if words v_a and v_b are related with respect to attributional information, while word v_b and v_c are related with respect to distributional information, that we might infer that v_a and v_c are related. Inferences may be made given the ensemble of correlations between and within the two data sources. Such inferences are suggested by the latent variables that are learned and shown in Figure 2. For example, in the case of the combined model's latent classes shown in Figure 2, we can see that attributional information leads to knowledge of patterns of body-parts, while distributional information leads to knowledge of patterns relating to medicine. Together both can lead to an inference that the realm of medicine and medical things are coupled with realm of body-parts and functions.

Conclusions and Further Work

Models that can integrate attributional and distributional information may inform us about how knowledge derived from experience in the world and knowledge derived from language may be integrated to lead to human semantic representations. This is the starting point and motivation behind this present work. Models based on attributional information develop semantic representations from the nonlinguistic attributes associated with the referents of words. This is an intuitively appealing idea about how word meanings are acquired. From a developmental perspective, these models provide an account of how word meanings can be linked to conceptual knowledge that develops independent of language. However, attributional information does not represent the only source of information about word meaning. From early stages in their development, children are exposed to language and it is reasonable to assume that they implicitly use the distributional information inherent in their linguistic input in order to learn new words, as well as to enrich the semantic representation for words referring to concrete referents. Using probabilistic models we have shown how these two different sources of information can be integrated to form semantic representations.

Two projects are to be carried out to further the ideas presented here. The first regards the models that are used. The EM algorithm is prone to local minima, as well as data over-fitting. This can limit the usefulness of the models that we have presented here. Bayesian learning methods based on MCMC sampling can be used to sample from the posterior probability of the parameters, and integrate over this for both learning and inference. These methods have already been pursued in related probabilistic models used for learning semantic representations (Blei & Ng, 2003; Griffiths & Steyvers, 2003, 2002; Teh et al., 2004). The second project concerns comparison with human behavioral data. We speculate that semantic relationships formed when both attributional and distributional information are used in combination will be more accurate in describing human performance than models based on either source alone.

References

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

Burgess, C., & Lund, K. (1997). Modeling parsing constraints with high-dimensional context-space. *Language and Cognitive Processes*, *12*, 177-210.

Collins, A., & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *12*, 240-247.

Dagan, I., Lee, L., & Pereira, F. (1997). Similarity-based methods for word sense disambiguation. In *35th annual meeting of the acl* (p. 56-63).

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis (special volume of the Philological Society, Oxford)* (p. 1-32). Oxford: Blackwell.

Griffiths, T., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the 24th annual conference of the cognitive science society.*

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.

Hinton, G., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*, 74-95.

Hofmann, T. (1999a). Probabilistic latent semantic analysis. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence.*

Hofmann, T. (1999b). Probabilistic latent semantic indexing. In *Proceedings of the twenty-second annual international sicir conference on research and development in information retrieval.*

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.

McClelland, J., & Rogers, T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310-322.

McRae, K., Sa, V. de, & Seidenberg, M. (1997). On the nature and scope of featural representation of word meaning. *Journal of Experimental Psychology: General*, *126*, 99-130.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), *The psychology of computer vision* (p. 211-277). McGraw-Hill.

Schütze, H. (1992). Dimensions of meaning. *Proceedings of Supercomputing.*

Smith, E., Shoben, E., & Rips, L. (1974). Structure and process in semantic memory: Featural model for semantic decisions. *Psyhological Review*, *81*, 214-241.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2004). Hierarchical dirichlet processes. In *Advances in neural information processing systems* (Vol. 17).

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, *48*, 422-488.