Probabilistic Distributional Semantics with Latent Variable Models

Diarmuid Ó Séaghdha*
University of Cambridge, UK

Anna Korhonen*
University of Cambridge, UK

We describe a probabilistic framework for acquiring selectional preferences of linguistic predicates and for using the acquired representations to model the effects of context on word meaning. Our framework uses Bayesian latent-variable models inspired by, and extending, the well-known Latent Dirichlet Allocation (LDA) model of topical structure in documents; when applied to predicate–argument data, topic models automatically induce semantic classes of arguments and assign each predicate a distribution over those classes. We consider LDA and a number of extensions to the model and evaluate them on a variety of semantic prediction tasks, demonstrating that our approach attains state-of-the-art performance. More generally, we argue that probabilistic methods provide an effective and flexible methodology for distributional semantics.

1. Introduction

Computational models of **lexical semantics** attempt to represent aspects of word meaning. For example, a model of the meaning of *dog* may capture the facts that dogs are animals, that they bark and chase cats, that they are often kept as pets, and so on. Word meaning is a fundamental component of the way language works: Sentences (and larger structures) consist of words, and their meaning is derived in part from the contributions of their constituent words' lexical meanings. At the same time, words instantiate a mapping between conceptual "world knowledge" and knowledge of language.

The relationship between the meanings of an individual word and the larger linguistic structure in which it appears is not unidirectional; while the word contributes to the meaning of the structure, the structure also clarifies the meaning of the word. Taken on its own a word may be vague or ambiguous, in the senses of Zwicky and Sadock (1975); even when the word's meaning is relatively clear it may still admit specification of additional details that affect its interpretation (e.g., what color/breed was the *dog*?). This specification comes through **context**, which consists of both linguistic and extralinguistic factors but shows a strong effect of the immediate lexical and syntactic environment—the other words surrounding the word of interest and their syntactic relations to it.

* 15 JJ Thomson Avenue, Cambridge, CB3 0FD, United Kingdom.
E-mail: Diarmuid.O'Seaghdha@cl.cam.ac.uk.

Submission received: 20 December 2012; revised version received: 14 July 2013; accepted for publication: 7 October 2013

doi:10.1162/COLI_a_00194

These diverse concerns motivate lexical semantic modeling as an important task for all computational systems that must tackle problems of meaning. In this article we develop a framework for modeling word meaning and how it is modulated by contextual effects.[1] Our models are **distributional** in the sense that their parameters are learned from observed co-occurrences between words and contexts in corpus data. More specifically, they are **probabilistic models** that associate latent variables with automatically induced classes of distributional behavior and associate each word with a probability distribution over those classes. This has a natural interpretation as a model of **selectional preference**, the semantic phenomenon by which predicates such as verbs or adjectives more plausibly combine with some classes of arguments than with others. It also has an interpretation as a disambiguation model: The different latent variable values correspond to different aspects of meaning and a word's distribution over those values can be modified by information coming from the context it appears in. We present a number of specific models within this framework and demonstrate that they can give state-of-the-art performance on tasks requiring models of preference and disambiguation. More generally, we illustrate that probabilistic modeling is an effective general-purpose framework for distributional semantics and a useful alternative to the popular vector-space framework.

The main contributions of the article are as follows:

- We describe the probabilistic approach to distributional semantics, showing how it can be applied as generally as the vector-space approach.
- We present three novel probabilistic selectional preference models and show that they outperform a variety of previously proposed models on a plausibility-based evaluation.
- Furthermore, the representations learned by these models correspond to semantic classes that are useful for modeling the effect of context on semantic similarity and disambiguation.

Section 2 presents background on distributional semantics and an overview of prior work on selectional preference learning and on modeling contextual effects. Section 3 introduces the probabilistic latent-variable approach and details the models we use. Section 4 presents our experimental results on four data sets. Section 5 concludes and sketches promising research directions for the future.

2. Background and Related Work

2.1 Distributional Semantics

The distributional approach to semantics is often traced back to the so-called "distributional hypothesis" put forward by mid-century linguists such as Zellig Harris and J.R. Frith:

> If we consider words or morphemes *A* and *B* to be more different in meaning than *A* and *C*, then we will often find that the distributions of *A* and *B* are more different than the distributions of *A* and *C*. (Harris 1954)

1 We build on previous work published in Ó Séaghdha (2010) and Ó Séaghdha and Korhonen (2011), adding new models and evaluation experiments as well as a comprehensive exposition. In Section 4 we indicate which experimental results have previously been reported.

You shall know a word by the company it keeps. (Frith 1957)

In Natural Language Processing (NLP), the term **distributional semantics** encompasses a broad range of methods that identify the semantic properties of a word or other linguistic unit with its patterns of co-occurrence in a corpus of textual data. The potential for learning semantic knowledge from text was recognized very early in the development of NLP (Spärck Jones 1964; Cordier 1965; Harper 1965), but it is with the technological developments of the past twenty years that this data-driven approach to semantics has become dominant. Distributional approaches may use a representation based on vector spaces, on graphs, or (like this article) on probabilistic models, but they all share the common property of estimating their parameters from empirically observed co-occurrences.

The basic unit of distributional semantics is the **co-occurrence**: an observation of a word appearing in a particular context. The definition is a general one: We may be interested in all kinds of words, or only a particular subset of the vocabulary; we may define the context of interest to be a document, a fixed-size window around a nearby word, or a syntactic dependency arc incident to a nearby word. Given a data set of co-occurrence observations we can extract an indexed set of co-occurrence counts $\mathbf{f}_w$ for each word of interest w; each entry f_{wc} counts the number of times that w was observed in context c. Alternatively, we can extract an indexed set $\mathbf{f}_c$ for each context.

The vector-space approach is the best-known methodology for distributional semantics; under this conception $\mathbf{f}_w$ is treated as a vector in $\mathbb{R}^{|\mathcal{C}|}$, where $\mathcal{C}$ is the vocabulary of contexts. As such, $\mathbf{f}_w$ is amenable to computations known from linear algebra. We can compare co-occurrence vectors for different words with a similarity function such as the cosine measure or a dissimilarity function such as Euclidean distance; we can cluster neighboring vectors; we can project a matrix of co-occurrence counts onto a low-dimensional subspace; and so on. This is perhaps the most popular approach to distributional semantics and there are many good general overviews covering the possibilities and applications of the vector space model (Curran 2003; Weeds and Weir 2005; Padó and Lapata 2007; Turney and Pantel 2010).

Although it is natural to view the aggregate of co-occurrence counts for a word as constituting a vector, it is equally natural to view it as defining a probability distribution. When normalized to have unit sum, $\mathbf{f}_w$ parameterizes a discrete distribution giving the conditional probability of observing a particular context given that we observe w. The contents of the vector-space modeler's toolkit generally have probabilistic analogs: similarity and dissimilarity can be computed using measures from information theory such as the Kullback–Leibler or Jensen–Shannon divergences (Lee 1999); the effects of clustering and dimensionality reduction can be achieved through the use of latent variable models (see Section 3.2.2). Additionally, Bayesian priors on parameter distributions provide a flexible toolbox for performing regularization and incorporating prior information in learning. A further advantage of the probabilistic framework is that it is often straightforward to extend existing models to account for additional structure in the data, or to tie together parameters for shared statistical strength, while maintaining guarantees of well-normalized behavior thanks to the laws of probability. In this article we focus on selectional preference learning and contextual disambiguation but we believe that the probabilistic approach exemplified here can fruitfully be applied in any scenario involving distributional semantic modeling.

2.2 Selectional Preferences

2.2.1 Motivation. A fundamental concept in linguistic knowledge is the **predicate**, by which we mean a word or other symbol that combines with one or more **arguments** to produce a composite representation with a composite meaning (by the principle of compositionality). The archetypal predicate is a verb; for example, transitive *drink* takes two noun arguments as subject and object, with which it combines to form a basic sentence. However, the concept is a general one, encompassing other word classes as well as more abstract items such as semantic relations (Yao et al. 2011), semantic frames (Erk, Padó, and Padó 2010), and inference rules (Pantel et al. 2007). The asymmetric distinction between predicate and argument is analogous to that between context and word in the more general distributional framework.

It is intuitive that a particular predicate will be more compatible with some semantic argument classes than with others. For example, the subject of *drink* is typically an animate entity (human or animal) and the object of *drink* is typically a beverage. The subject of *eat* is also typically an animate entity but its object is typically a foodstuff. The noun modified by the adjective *tasty* is also typically a foodstuff, whereas the noun modified by *informative* is an information-bearing object. This intuition can be formalized in terms of a predicate's **selectional preference**: a function that assigns a numerical score to a combination of a predicate and one or more arguments according to the semantic plausibility of that combination. This score may be a probability, a rank, a real value, or a binary value; in the last case, the usual term is **selectional restriction**.

Models of selectional preference aim to capture conceptual knowledge that all language users are assumed to have. Speakers of English can readily identify that examples such as the following are semantically infelicitous despite being syntactically well-formed:

1. The beer drank the man.

2. Quadruplicity drinks procrastination. (Russell 1940)

3. Colorless green ideas sleep furiously. (Chomsky 1957)

4. The paint is silent. (Katz and Fodor 1963)

Psycholinguistic experiments have shown that the time course of human sentence processing is sensitive to predicate–argument plausibility (Altmann and Kamide 1999; Rayner et al. 2004; Bicknell et al. 2010): Reading times are faster when participants are presented with plausible combinations than when they are presented with implausible combinations. It has also been proposed that selectional preference violations are cues that trigger metaphorical interpretation. Wilks (1978) gives the example *My car drinks gasoline*, which must be understood non-literally since *car* strongly violates the subject preference of *drink* and *gasoline* is also an unlikely candidate for something to drink.

In NLP, one motivation for modeling predicate–argument plausibility is to investigate whether this aspect of human conceptual knowledge can be learned automatically from text corpora. If the predictions of a computational model correlate with judgments collected from human behavioral data, the assumption is that the model itself shares some properties with human linguistic knowledge and is in some sense a "good" semantic model. More practically, NLP researchers have shown that selectional preference knowledge is useful for downstream applications, including metaphor detection (Shutova 2010), identification of non-compositional multiword

expressions (McCarthy, Venkatapathy, and Joshi 2007), semantic role labeling (Gildea and Jurafsky 2002; Zapirain, Agirre, and Màrquez 2009; Zapirain et al. 2010), word sense disambiguation (McCarthy and Carroll 2003), and parsing (Zhou et al. 2011).

2.2.2 The "Counting" Approach. The simplest way to estimate the plausibility of a predicate–argument combination from a corpus is to count the number of times that combination appears, on the assumptions that frequency correlates with plausibility and that given enough data the resulting estimates will be relatively accurate. For example, Keller and Lapata (2003) estimate predicate–argument plausibilities by submitting appropriate queries to a Web search engine and counting the number of "hits" returned. To estimate the frequency with which the verb *drink* takes *beer* as a direct object, Keller and Lapata's method uses the query *<drink|drinks|drank|drunk|drinking a|the|∅ beer|beers>*; to estimate the frequency with which *tasty* modifies *pizza* the query is simply *<tasty pizza|pizzas>*. Where desired, these joint frequency counts can be normalized by unigram hit counts to estimate conditional probabilities such as $P(pizza|tasty)$.

The main advantages of this approach are its simplicity and its ability to exploit massive corpora of raw text. On the other hand, it is hindered by the facts that only shallow processing is possible and that even in a Web-scale corpus the probability estimates for rare combinations will not be accurate. At the time of writing, Google returns zero hits for the query *<draughtsman|draughtsmen whistle|whistles|whistled|whistling>* and 1,570 hits for *<onion|onions whistle|whistles|whistled|whistling>*, suggesting the implausible conclusion that an onion is far more likely to whistle than a draughtsman.[2]

Zhou et al. (2011) modify the Web query approach to better capture statistical association by using pointwise mutual information (PMI) rather than raw co-occurrence frequency to quantify selectional preference:

$$PMI(p, a) = \log \frac{P(p, a)}{P(p)P(a)} \tag{1}$$

The role of the PMI transformation is to correct for the effect of unigram frequency: A common word may co-occur often with another word just because it is a common word rather than because there is a semantic association between them. However, it does not provide a way to overcome the problem of inaccurate counts for low-probability co-occurrences. Zhou et al.'s goal is to incorporate selectional preference features into a parsing model and they do not perform any evaluation of the semantic quality of the resulting predictions.

2.2.3 Similarity-Based Smoothing Methods. During the 1990s, research on language modeling led to the development of various "smoothing" methods for overcoming the data sparsity problem that inevitably arises when estimating co-occurrence counts from finite corpora (Chen and Goodman 1999). The general goal of smoothing algorithms is to alter the distributional profile of observed counts to better match the known statistical properties of linguistic data (e.g., that language exhibits power-law behavior). Some also incorporate semantic information on the assumption that meaning guides the distribution of words in a text.

2 The analogous example given by Ó Séaghdha (2010) relates to the plausibility of a manservant or a carrot laughing; Google no longer returns zero hits for *<a|the manservant|manservants|menservants laugh|laughs|laughed>* but a frequency-based estimate still puts the probability of a carrot laughing at 200 times that of a manservant laughing (1,680 hits against 81 hits).

One such class of methods is based on **similarity-based smoothing**, by which one can extrapolate from observed co-occurrences by implementing the distributional hypothesis: "similar" words will have similar distributional properties. A general form for similarity-based co-occurrence estimates is

$$P(w_2|w_1) = \sum_{w_3 \in \mathcal{S}(w_1,w_2)} \frac{sim(w_2,w_3)}{\sum_{w' \in \mathcal{S}(w_1,w_2)} sim(w_2,w')} P(w_3|w_1) \quad (2)$$

sim can be an arbitrarily chosen similarity function; Dagan, Lee, and Pereira (1999) investigate a number of options. $\mathcal{S}(w_1, w_2)$ is a set of comparison words that may depend on w_1 or w_2, or neither: Essen and Steinbiss (1992) use the entire vocabulary, whereas Dagan, Lee, and Pereira use a fixed number of the most similar words to w_2, provided their similarity value is above a threshold t.

While originally proposed for language modeling—the task of estimating the probability of a sequence of words—these methods require only trivial alteration to estimate co-occurrence probabilities for predicates and arguments, as was noted early on by Grishman and Sterling (1993) and Dagan, Lee, and Pereira (1999). Erk (2007) and Erk, Padó, and Padó (2010) build on this prior work to develop an "exemplar-based" selectional preference model called EPP. In the EPP model, the set of comparison words is the set of words observed for the predicate p in the training corpus, denoted $Seenargs(p)$:

$$Selpref_{EPP}(a|p) = \sum_{a' \in Seenargs(p)} \frac{weight(a'|p) sim(a',a)}{\sum_{a'' \in Seenargs(p)} weight(a''|p)} \quad (3)$$

The co-occurrence strength $weight(a|p)$ may simply be normalized co-occurrence frequency; alternatively a statistical association measure such as pointwise mutual information may be used. As before, $sim(a, a')$ may be any similarity measure defined on members of A. One advantage of this and other similarity-based models is that the corpus used to estimate similarity need not be the same as that used to estimate predicate–argument co-occurrence, which is useful when the corpus labeled with these co-occurrences is small (e.g., a corpus labeled with FrameNet frames).

2.2.4 Discriminative Models. Bergsma, Lin, and Goebel (2008) cast selectional preference acquisition as a supervised learning problem to which a discriminatively trained classifier such as a Support Vector Machine (SVM) can be applied. To produce training data for a predicate, they pair "positive" arguments that were observed for that predicate in the training corpus and have an association with that predicate above a specified threshold (measured by mutual information) with randomly selected "negative" arguments of similar frequency that do not occur with the predicate or fall below the association threshold. Given this training data, a classifier can be trained in a standard way to predict a positive or negative score for unseen predicate–argument pairs.

An advantage of this approach is that arbitrary sets of features can be used to represent the training and testing items. Bergsma, Lin, and Goebel include conditional probabilities $P(a|p)$ for all predicates the candidate argument co-occurs with, typographic features of the argument itself (e.g., whether it is capitalized, or contains digits), lists of named entities, and precompiled semantic classes.

2.2.5 WordNet-Based Models. An alternative approach to preference learning models the argument distribution for a predicate as a distribution over semantic classes provided by a predefined lexical resource. The most popular such resource is the WordNet lexical hierarchy (Fellbaum 1998), which provides semantic classes and hypernymic structures for nouns, verbs, adjectives, and adverbs.[3] Incorporating knowledge about the WordNet taxonomy structure in a preference model enables the use of graph-based regularization techniques to complement distributional information, while also expanding the coverage of the model to types that are not encountered in the training corpus. On the other hand, taxonomy-based methods build in an assumption that the lexical hierarchy chosen is the universally "correct" one and they will not perform as well when faced with data that violates the hierarchy or contains unknown words. A further issue faced by these models is that the resources they rely on require significant effort to create and will not always be available to model data in a new language or a new domain.

Resnik (1993) proposes a measure of associational strength between a predicate and WordNet classes based on the empirical distribution of words of each class (and their hyponyms) in a corpus. Abney and Light (1999) conceptualize the process of generating an argument for a predicate in terms of a Markovian random walk from the hierarchy's root to a leaf node and choosing the word associated with that leaf node. Ciaramita and Johnson (2000) likewise treat WordNet as defining the structure of a probabilistic graphical model, in this case a Bayesian network. Li and Abe (1998) and Clark and Weir (2002) both describe models in which a predicate "cuts" the hierarchy at an appropriate level of generalization, such that all classes below the cut are considered appropriate arguments (whether observed in data or not) and all classes above the cut are considered inappropriate.

In this article we focus on purely distributional models that do not rely on manually constructed lexical resources; therefore we do not revisit the models described in this section subsequently, except as a basis for empirical comparison. Ó Séaghdha and Korhonen (2012) do investigate a number of Bayesian preference models that incorporate WordNet classes and structure, finding that such models outperform previously proposed WordNet-based models and perform comparably to the distributional Bayesian models presented here.

2.3 Measuring Similarity in Context

2.3.1 Motivation. A fundamental idea in semantics is that the meaning of a word is disambiguated and modulated by the context in which it appears. The word *body* clearly has a different sense in each of the following text fragments:

1. *Depending on the present position of the planetary body in its orbital path, . . .*
2. *The executive body decided. . .*
3. *The human body is intriguing in all its forms.*

In a standard word sense disambiguation experiment, the task is to map instances of ambiguous words onto senses from a manually compiled inventory such as WordNet. An alternative experimental method is to have a system rate the suitability of replacing an ambiguous word with an alternative word that is synonymous or semantically

3 WordNet also contains many other kinds of semantic relations besides hypernymy but these are not typically used for selectional preference modeling.

similar in some contexts but not others. For example, *committee* is a reasonable substitute for *body* in fragment 2 but less reasonable in fragment 1. An evaluation of semantic models based on this principle was run as the English Lexical Substitution Task in SemEval 2007 (McCarthy and Navigli 2009). The annotated data from the Lexical Substitution Task have been used by numerous researchers to evaluate models of lexical choice; see Section 4.5 for further details.

In this section we formalize the problem of predicting the similarity or substitutability of a pair of words w_o, w_s in a given context $C = \{(r^1, w^1), (r^2, w^2), \ldots, (r^n, w^n)\}$. When the task is substitution, w_o is the original word and w_s is the candidate substitute. Our general approach is to compute a representation $Rep(w_o|C)$ for w_o in context C and compare it with $Rep(w_s)$, our representation for w_n:

$$sim(w_o, w_s|C) = sim(Rep(w_o|C), Rep(w_s)) \tag{4}$$

where *sim* is a suitable similarity function for comparing the representations. This general framework leaves open the question of what kind of representation we use for $Rep(w_o|C)$ and $Rep(w_s)$; in Section 2.3.2 we describe representations based on vector-space semantics and in Section 3.5 we describe representations based on latent-variable models.

A complementary perspective on the disambiguatory power of context models is provided by research on semantic composition, namely, how the syntactic effect of a grammar rule is accompanied by a combinatory semantic effect. In this view, the goal is to represent the combination of a context and an in-context word, not just to represent the word given the context. The co-occurrence models described in this article are not designed to scale up and provide a representation for complex syntactic structures,[4] but they are applicable to evaluation scenarios that involve representing binary co-occurrences.

2.3.2 Vector-Space Models. As described in Section 2.1, the vector-space approach to distributional semantics casts word meanings as vectors of real numbers and uses linear algebra operations to compare and combine these vectors. A word w is represented by a vector $\mathbf{v_w}$ that models aspects of its distribution in the training corpus; the elements of this vector may be co-occurrence counts (in which case it is the same as the frequency vector $\mathbf{f_w}$) or, more typically, some transformation of the raw counts.

Mitchell and Lapata (2008, 2010) present a very general vector-space framework in which to consider the problem of combining the semantic representations of co-occurring words. Given pre-computed word vectors $\mathbf{v_w}$, $\mathbf{v_{w'}}$, their combination $\mathbf{p}$ is provided by a function g that may also depend on syntax R and background knowledge K:

$$\mathbf{p} = g(\mathbf{v_w}, \mathbf{v_{w'}}, R, K) \tag{5}$$

Mitchell and Lapata investigate a number of functions that instantiate Equation (5), finding that elementwise multiplication is a simple and consistently effective choice:

$$p_i = v_{wi} \cdot v_{w'i} \tag{6}$$

4 cf. Grefenstette and Sadrzadeh (2011), Socher et al. (2011)

The motivation for this "disambiguation by multiplication" is that lexical vectors are sparse and the multiplication operation has the effect of sending entries not supported in both $\mathbf{v_w}$ and $\mathbf{v_{w'}}$ towards zero while boosting entries that have high weights in both vectors.

The elementwise multiplication approach assumes that all word vectors are in the same space. For a syntactic co-occurrence model, this is often not the case: The contexts for a verb and a noun may have no dependency labels in common and hence multiplying their vectors will not give useful results. Erk and Padó (2008) propose a **structured vector space** approach in which each word w is associated with a set of "expectation" vectors R_w, indexed by dependency label, in addition to its standard co-occurrence vector $\mathbf{v_w}$. The expectation vector $R_w(r)$ for word w and dependency label r is an average over co-occurrence vectors for seen arguments of w and r in the training corpus:

$$R_w(r) = \sum_{w' : f(w,r,w') > 0} f(w, r, w') \cdot \mathbf{v_{w'}} \tag{7}$$

Whereas a standard selectional preference model addresses the question "which words are probable as arguments of predicate (w, r)?", the expectation vector (7) addresses the question "what does a typical co-occurrence vector for an argument of the predicate (w, r) look like?". To disambiguate the semantics of word w in the context of a predicate (w', r'), Erk and Padó combine the expectation vector $R_{w'}(r')$ with the word vector $\mathbf{v_w}$:

$$\mathbf{v_{w|r',w'}} = R_{w'}(r') \cdot \mathbf{v_w} \tag{8}$$

Thater, Fürstenau, and Pinkal (2010, 2011) have built on the idea of using syntactic vector spaces for disambiguation. The model of Thater, Fürstenau, and Pinkal (2011), which is simpler and better-performing, sets the representation of w in the context of (r', w') to be

$$\mathbf{v_{w|r',w'}} = \sum_{w'',r''} \alpha_{w',r',w'',r''} \cdot weight(w'', r'', w) \cdot e_{r'',w''} \tag{9}$$

where α quantifies the compatibility of the observed predicate (w', r') with the smoothing predicate (w'', r''), *weight* quantifies the co-occurrence strength between (w'', r'') and w, and $e_{r'',w''}$ is a basis vector for (w'', r''). This is a general formulation admitting various choices of α and *weight*; the optimal configuration is found to be as follows:

$$\alpha_{w',r',w'',r''} = \begin{cases} sim(\mathbf{v_{w'}}, \mathbf{v_{w''}}) & \text{if } r' = r'' \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$$weight(w'', r'', w) = PMI((w'', r''), w) \tag{11}$$

This is conceptually very similar to the EPP selectional preference model (3) of Erk, Padó, and Padó (2010); each entry in the vector $\mathbf{v_{w|r',w'}}$ is a similarity-smoothed estimate of the preference of (w', r') for w. EPP uses seen arguments of (w', r') for smoothing, whereas Thater, Fürstenau, and Pinkal (2011) take a complementary approach and

smooth with seen predicates for w. In order to combine the disambiguatory effects of multiple predicates, a sum over contextualized vectors is taken:

$$\mathbf{v}_{\mathbf{w}|(\mathbf{r^1},\mathbf{w^1}),(\mathbf{r^2},\mathbf{w^2}),\ldots,(\mathbf{r^n},\mathbf{w^n})} = \sum_{i}^{n} \mathbf{v}_{\mathbf{w}|\mathbf{r^i},\mathbf{w^i}} \tag{12}$$

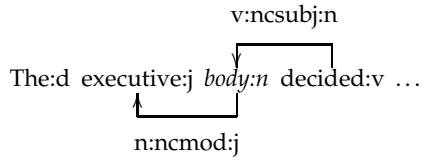
All the models described in this section provide a way of relating a word's standard co-occurrence vector to a vector representation of the word's meaning in context. This allows us to calculate the similarity between two in-context words or between a word and an in-context word using standard vector similarity measures such as the cosine. In applications where the task is to judge the appropriateness of substituting a word w_s for an observed word w_o in context $C = \{(r^1, w^1), (r^2, w^2), \ldots, (r^n, w^n)\}$, a common approach is to compute the similarity between the contextualized vector $\mathbf{v}_{\mathbf{w_o}|(\mathbf{r^1},\mathbf{w^1}),(\mathbf{r^2},\mathbf{w^2}),\ldots,(\mathbf{r^n},\mathbf{w^n})}$ and the uncontextualized word vector v_{w_s}. It has been demonstrated empirically that this approach yields better performance than contextualizing both vectors before the similarity computation.

3. Probabilistic Latent Variable Models for Lexical Semantics

3.1 Notation and Terminology

We define a co-occurrence as a pair (c, w), where c is a context belonging to the vocabulary of contexts $\mathcal{C}$ and w is a word belonging to the word vocabulary $\mathcal{W}$.[5] Unless otherwise stated, the contexts considered in this article are head-lexicalized dependency edges $c = (r, w_h)$ where $r \in \mathcal{R}$ is the grammatical relation and $w_h \in \mathcal{W}$ is the head lemma. We notate grammatical relations as p_h:*label*:p_d, where p_h is the head word's part of speech, p_d is the dependent word's part of speech, and *label* is the dependency label.[6] We use a coarse set of part-of-speech tags: n (noun), v (verb), j (adjective), r (adverb). The dependency labels are the grammatical relations used by the RASP system (Briscoe 2006; Briscoe, Carroll, and Watson 2006), though in principle any dependency formalism could be used. The assumption that predicates correspond to head-lexicalized dependency edges means that they have arity one.

Given a parsed sentence, each word w in the sentence has a syntactic context set C comprising all the dependency edges incident to w. In the sentence fragment *The executive body decided...*, the word *body* has two incident edges:



5 When specifically discussing selectional preferences, we will also use the terms **predicate** and **argument** to describe a co-occurrence pair; when restricted to syntactic predicates, the former term is synonymous with our definition of context.

6 Strictly speaking, w and w_h are drawn from subsets of $\mathcal{W}$ that are licensed by r when r is a syntactic relation, that is, they must have parts of speech p_d and p_h, respectively. Our models assume a fixed argument vocabulary, so we can partition the training data according to part of speech; the models are agnostic regarding the predicate vocabulary as these are subsumed by the context vocabulary. In the interest of parsimony we leave this detail implicit in our notation.

The context set for *body* is $C = \{(j{:}ncmod^{-1}{:}n{,}executive), (v{:}ncsubj{:}n{,}decide)\}$, where (*v:ncsubj:n,decide*) indicates that *body* is the subject of *decide* and (*j:ncmod*$^{-1}$*:n,executive*) denotes that it stands in an inverse non-clausal modifier relation to *executive* (we assume that nouns are the heads of their adjectival modifiers).

To estimate our preference models we will rely on co-occurrence counts extracted from a corpus of observations O. Each observation is a co-occurrence of a predicate and an argument. The set of observations for context c is denoted $O(c)$. The co-occurrence frequency of context c and word w is denoted by f_{cw}, and the total co-occurrence frequency of c by $f_c = \sum_{w \in W} f_{cw}$.

3.2 Modeling Assumptions

3.2.1 Bayesian Modeling. The Bayesian approach to probabilistic modeling (Gelman et al. 2003) is characterized by (1) the use of prior distributions over model parameters to encode the modeler's expectations about the values they will take; and (2) the explicit quantification of uncertainty by maintaining posterior distributions over parameters rather than point estimates.[7]

As is common in NLP, the data we are interested in modeling are drawn from a discrete sample space (e.g., the vocabulary of words or a set of semantic classes). This leads to the use of a categorical or multinomial distribution for the data likelihood. This distribution is parameterized by a unit-sum vector $\boldsymbol{\theta}$ with length $|K|$ where K is the sample space. The probability that an observation o takes value k is then:

$$o \sim \text{Multinomial}(\boldsymbol{\theta}) \tag{13}$$

$$P(o = k|\boldsymbol{\theta}) = \theta_k \tag{14}$$

The value of $\boldsymbol{\theta}$ must typically be learned from data. The maximum likelihood estimate (MLE) sets θ_k proportional to the number of times k was observed in a set of observations O, where each observation $o_i \in K$:

$$\theta_k^{MLE} = \frac{f_k}{|O|} \tag{15}$$

Although simple, such an approach has significant limitations. Because a linguistic vocabulary contains a large number of items that individually have low probability but together account for considerable total probability mass, even a large corpus is unlikely to give accurate estimates for low-probability types (Evert 2004). Items that do not appear in the training data will be assigned zero probability of appearing in unseen data, which is rarely if ever a valid assumption. Sparsity increases further when the sample space contains composite items (e.g., context-words pairs).

The standard approach to dealing with the shortcomings of MLE estimation in language modeling is to "smooth" the distribution by taking probability mass from frequent types and giving it to infrequent types. The Bayesian approach to smoothing is to place an appropriate prior on $\boldsymbol{\theta}$ and apply Bayes' Theorem:

$$P(\boldsymbol{\theta}|O) = \frac{P(O|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int P(O|\boldsymbol{\theta})P(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \tag{16}$$

7 However, the second point is often relaxed in application contexts where the posterior mean is used for inference (e.g., Section 3.4.2).

A standard choice for the prior distribution over the parameters of a discrete distribution is the Dirichlet distribution:

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha}) \tag{17}$$

$$P(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1} \tag{18}$$

Here, $\boldsymbol{\alpha}$ is a $|K|$-length vector where each $\alpha_k > 0$. One effect of the Dirichlet prior is that setting the sum $\sum_k \alpha_k$ to a small value will encode the expectation that the parameter vector $\boldsymbol{\theta}$ is likely to distribute its mass more sparsely. The Dirichlet distribution is a **conjugate prior** for multinomial and categorical likelihoods, in the sense that the posterior distribution $P(\boldsymbol{\theta}|O)$ in Equation (16) is also a Dirichlet distribution when $P(O|\boldsymbol{\theta})$ is multinomial or categorical and $P(\boldsymbol{\theta})$ is Dirichlet:

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\mathbf{f}_O \oplus \boldsymbol{\alpha}) \tag{19}$$

where $\oplus$ indicates elementwise addition of the observed count vector $\mathbf{f}_O$ to the Dirichlet parameter vector $\boldsymbol{\alpha}$. Furthermore, the conjugacy property allows us to do a number of important computations in an efficient way. In many applications we are interested in predicting the distribution over values K for a "new" observation given a set of prior observations O while retaining our uncertainty about the model parameters. We can average over possible values of $\boldsymbol{\theta}$, weighted according to their probability $P(\boldsymbol{\theta}|O, \alpha)$ by "integrating out" the parameter and still retain a simple closed-form expression for the posterior predictive distribution:

$$P(o_{|O|+1} = k|O, \alpha) = \int P(o_{|O|+1} = k|\boldsymbol{\theta}) P(\boldsymbol{\theta}|O, \alpha) \mathrm{d}\boldsymbol{\theta} \tag{20}$$

$$= \frac{f_k + \alpha_k}{|O| + \sum_{k'} \alpha_{k'}} \tag{21}$$

Expression (21) is central to the implementation of collapsed Gibbs samplers for Bayesian models such as latent Dirichlet allocation (Section 3.3). For mathematical details of these derivations, see Heinrich (2009).

Other priors commonly used for discrete distributions in NLP include the Dirichlet process and the Pitman–Yor process (Goldwater, Griffiths, and Johnson 2011). The Dirichlet process provides similar behavior to the Dirichlet distribution prior but is "non-parametric" in the sense of varying the size of its support according to the data; in the context of mixture modeling, a Dirichlet process prior allows the number of mixture components to be learned rather than fixed in advance. The Pitman–Yor process is a generalization of the Dirichlet process that is better suited to learning power-law distributions. This makes it particularly suitable for language modeling where the Dirichlet distribution or Dirichlet process would not produce a long enough tail due to their preference for sparsity (Teh 2006). On the other hand, Dirichlet-like behavior may be preferable in semantic modeling, where we expect, for example, predicate–class and class–argument distributions to be sparse.

3.2.2 The Latent Variable Assumption. In probabilistic modeling, **latent variables** are random variables whose values are not provided by the input data. As a result, their

values must be inferred at the same time as the model parameters on the basis of the training data and model structure. The latent variable concept is a very general one that is used across a wide range of probabilistic frameworks, from hidden Markov models to neural networks. One important application is in mixture models, where the data likelihood is assumed to have the following form:

$$P(x) = \sum_z P(x|z)P(z) \tag{22}$$

Here the latent variables z index mixture components, each of which is associated with a distribution over observations x, and the resulting likelihood is an average of the component distributions weighted by the mixing weights $P(z)$. The set of possible values for z is the set of components Z. When $|Z|$ is small relative to the size of the training data, this model has a clustering effect in the sense that the distribution learned for $P(x|z)$ is informed by all datapoints assigned to component z.

In a model of two-way co-occurrences each observation consists of two discrete variables c and w, drawn from vocabularies $\mathcal{C}$ and $\mathcal{W}$, respectively.

$$P(w|c) = \sum_z P(w|z)P(z|c) \tag{23}$$

The idea of compressing the observed co-occurrence data through a small layer of latent variables shares the same basic motivations as other, not necessarily probabilistic, dimensionality reduction techniques such as Latent Semantic Analysis or Non-negative Matrix Factorization. An advantage of probabilistic models is their flexibility, both in terms of learning methods and model structures. For example, the models considered in this article can potentially be extended to multi-way co-occurrences and to hierarchically defined contexts that cannot easily be expressed in frameworks that require the input to be a $|\mathcal{C}| \times |\mathcal{W}|$ co-occurrence matrix.

To the best of our knowledge, latent variable models were first applied to co-occurrence data in the context of noun clustering by Pereira, Tishby, and Lee (1993). They suggest a factorization of a noun n's distribution over verbs v as

$$P(v|n) = \sum_z P(v|z)P(z|n) \tag{24}$$

which is equivalent to Equation (23) when we take n as the predicate and v as the argument, in effect defining an inverse selectional preference model. Pereira, Tishby, & Lee also observe that given certain assumptions Equation (24) can be written more symmetrically as

$$P(v,n) = \sum_z P(v|z)P(n|z)P(z) \tag{25}$$

The distributions $P(v|z)$, $P(n|z)$, and $P(z)$ are estimated by an optimization procedure based on Maximum Entropy. Rooth et al. (1999) propose a much simpler Expectation Maximization (EM) procedure for estimating the parameters of Equation (25).

3.3 Bayesian Models for Binary Co-occurrences

Combining the latent variable co-occurrence model (23) with the use of Dirichlet priors naturally leads to **Latent Dirichlet Allocation** (LDA) (Blei, Ng, and Jordan 2003). Often described as a "topic model," LDA is a model of document content that assumes each document is generated from a mixture of multinomial distributions or "topics." Topics are shared across documents and correspond to thematically coherent patterns of word usage. For example, one topic may assign high probability to the words *finance*, *fund*, *bank*, and *invest*, whereas another topic may assign high probability to the words *football*, *goal*, *referee*, and *header*. LDA has proven to be a very successful model with many applications and extensions, and the topic modeling framework remains an area of active research in machine learning.

Although originally conceived for modeling document content, LDA can be applied to any kind of discrete binary co-occurrence data. The original application of LDA is essentially a latent-variable model of document–word co-occurrence. Adapting LDA for selectional preference modeling was suggested independently by Ó Séaghdha (2010) and Ritter, Mausam, and Etzioni (2010). Conceptually the shift is straightforward and intuitive: Documents become contexts and words become argument words. The selectional preference probability $P(w|c)$ is modeled as

$$P(w|c) = \sum_z P(z|c)P(w|z) \tag{26}$$

Figure 1 sketches the "generative story" according to which LDA generates arguments for predicates and also presents a plate diagram indicating the dependencies between variables in the model. Table 1 illustrates the semantic representation induced by a 600-topic LDA model trained on predicate–noun co-occurrences extracted from the British National Corpus (for more details of this training data, see Section 4.1). The "semantic classes" are actually distributions over all nouns in the vocabulary rather than a hard partitioning; therefore we present the eight most probable words for each. We also present the contexts most frequently associated with each class. Whereas a

for topic $z \in \{1 \dots |Z|\}$ **do**
  (Draw a distribution over words)
  $\boldsymbol{\Phi}_z \sim Dirichlet(\boldsymbol{\beta})$
end for
for context $c \in \{1 \dots |\mathcal{C}|\}$ **do**
  (Draw a distribution over classes)
  $\boldsymbol{\theta}_c \sim Dirichlet(\boldsymbol{\alpha})$
  for observation $o_i \in O(c)$ **do**
    (Draw a class)
    $z_i \sim Multinomial(\boldsymbol{\theta}_c)$
    (Draw a word)
    $w_i \sim Multinomial(\boldsymbol{\Phi}_{z_i})$
  end for
end for

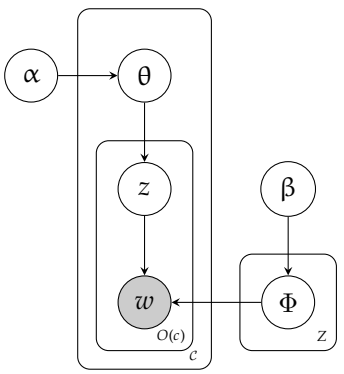


Figure 1
Generative story and plate diagram for LDA; descriptive comments (in parentheses) precede each sampling step.

Table 1
Sample semantic classes learned by an LDA syntactic co-occurrence model with $|Z| = 600$ trained on BNC co-occurrences.

Class 1	**Words:** attack, raid, assault, campaign, operation, incident, bombing **Object of:** launch, carry, follow, suffer, lead, mount, plan, condemn **Subject of:** happen, come, begin, cause, continue, take, follow **Modifies:** raid, furnace, shelter, victim, rifle, warning, aircraft **Modified by:** heart, bomb, air, terrorist, indecent, latest, further, bombing **Prepositional:** on home, on house, by force, target for, hospital after, die after
Class 2	**Words:** line, axis, section, circle, path, track, arrow, curve **Object of:** draw, follow, cross, dot, break, trace, use, build, cut **Subject of:** divide, run, represent, follow, indicate, show, join, connect **Modifies:** manager, number, drawing, management, element, treatment **Modified by:** straight, railway, long, cell, main, front, production, product **Prepositional:** on map, by line, for year, line by, point on, in fig, angle to
Class 3	**Words:** test, examination, check, testing, exam, scan, assessment, sample **Object of:** pass, carry, use, fail, perform, make, sit, write, apply **Subject of:** show, reveal, confirm, prove, consist, come, take, detect, provide **Modifies:** result, examination, score, case, ban, question, board, paper, kit **Modified by:** blood, medical, final, routine, breath, fitness, driving, beta **Prepositional:** subject to, at end, success in, on part, performance on
Class 4	**Words:** university, college, school, polytechnic, institute, institution, library **Object of:** enter, attend, leave, visit, become, found, involve, close, grant **Subject of:** offer, study, make, become, develop, win, establish, undertake **Modifies:** college, student, library, course, degree, department, school **Modified by:** university, open, technical, city, education, state, technology **Prepositional:** student at, course at, study at, lecture at, year at
Class 5	**Words:** fund, reserve, eyebrow, revenue, awareness, conservation, alarm **Object of:** raise, set, use, provide, establish, allocate, administer, create **Subject of:** raise, rise, shoot, lift, help, remain, set, cover, hold **Modifies:** manager, asset, raiser, statement, management, commissioner **Modified by:** nature, pension, international, monetary, national, social, trust **Prepositional:** for nature, contribution to, for investment, for development

topic model trained on document–word co-occurrences will find topics that reflect broad thematic commonalities, the model trained on syntactic co-occurrences finds semantic classes that capture a much tighter sense of similarity: Words assigned high probability in the same topic tend to refer to entities that have similar properties, that perform similar actions, and have similar actions performed on them. Thus Class 1 is represented by *attack*, *raid*, *assault*, *campaign,* and so on, forming a coherent semantic grouping. Classes 2, 3, and 4 correspond to groups of tests, geometric objects, and public/educational institutions, respectively. Class 5 has been selected to illustrate a potential pitfall of using syntactic co-occurrences for semantic class induction: *fund*, *revenue*, *eyebrow,* and *awareness* hardly belong together as a coherent conceptual class. The reason, it seems, is that they are all entities that can be (and in the corpus, are) *raised*. This class has also conflated different (but related) senses of *reserve* and as a result the modifier *nature* is often associated with it.

An alternative approach is suggested by the model used by Pereira, Tishby, and Lee (1993) and Rooth et al. (1999) that is formalized in Equation (25). This model can

(Draw a distribution over topics)
$\boldsymbol{\theta} \sim Dirichlet(\boldsymbol{\alpha})$
for topic $z \in \{1 \ldots |Z|\}$ **do**
  (Draw a distribution over words)
  $\boldsymbol{\Phi}_z \sim Dirichlet(\boldsymbol{\beta})$
  (Draw a distribution over contexts)
  $\boldsymbol{\Psi}_z \sim Dirichlet(\boldsymbol{\gamma})$
end for
for observation $o_i \in O$ **do**
  (Draw a topic)
  $z_i \sim Multinomial(\boldsymbol{\theta})$
  (Draw a word)
  $w_i \sim Multinomial(\boldsymbol{\Phi}_{z_i})$
  (Draw a context)
  $c_i \sim Multinomial(\boldsymbol{\Psi}_{z_i})$
end for

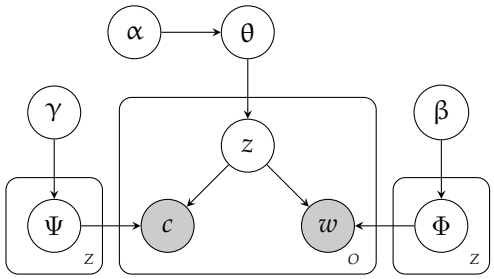


Figure 2
Generative story and plate diagram for ROOTH-LDA.

be "Bayesianized" by placing Dirichlet priors on the component distributions; adapting Equation (25) to our notation, the resulting joint distribution over contexts and words is

$$P(c, w) = \sum_z P(c|z)P(w|z)P(z) \tag{27}$$

The generative story and plate diagram for this model, which was called ROOTH-LDA in Ó Séaghdha (2010), are given in Figure 2. Whereas LDA induces classes of arguments, ROOTH-LDA induces classes of predicate–argument interactions. Table 2 illustrates some classes learned by ROOTH-LDA from BNC verb–object co-occurrences. One class shows that a *cost*, *number*, *risk,* or *expenditure* can plausibly be *increased*, *reduced*, *cut*, or *involved*; another shows that a *house*, *building*, *home,* or *station* can be *built*, *left*, *visited,* or *used*. As with LDA, there are some over-generalizations; the fact that an *eye* or *mouth* can be *opened*, *closed,* or *shut* does not necessarily entail that it can be *locked* or *unlocked*.

For many predicates, the best description of their argument distributions is one that accounts for general semantic regularities and idiosyncratic lexical patterns. This suggests the idea of combining a distribution over semantic classes and a predicate-specific

Table 2
Sample semantic classes learned by a Rooth-LDA model with $|Z| = 100$ trained on BNC verb–object co-occurrences.

Class 1		Class 2		Class 3		Class 4	
increase	cost	open	door	build	house	spend	time
reduce	number	close	eye	leave	building	work	day
cut	risk	shut	mouth	visit	home	wait	year
involve	expenditure	lock	window	use	station	come	hour
control	demand	slam	gate	enter	church	waste	night
estimate	pressure	unlock	shop	include	school	take	week
limit	rate	keep	fire	see	plant	remember	month
cover	power	round	book	run	office	end	life

distribution over arguments. One way of doing this is through the model depicted in Figure 3, which we call LEX-LDA; this model defines the selectional preference probability $P(w|c)$ as

$$P(w|c) = \sigma_c P_{lex}(w|c) + (1 - \sigma_c) P_{class}(w|c) \tag{28}$$

$$= \sigma_c P_{lex}(w|c) + (1 - \sigma_c) \sum_z P(w|z) P(z|c) \tag{29}$$

where σ_c is a value between 0 and 1 that can be interpreted as a measure of argument lexicalization or as the probability that an observation for context c is drawn from the lexical distribution P_{lex} or the class-based distribution P_{class}. P_{class} has the same form as the LDA preference model. The value of σ_c will vary across predicates according to how well their argument preference can be fit by the class-based models; a predicate with high σ_c will have idiosyncratic argument patterns that are best learned by observing that predicate's co-occurrences in isolation. In many cases this may reflect idiomatic or non-compositional usages, though it is also to be expected that σ_c will correlate with frequency; given sufficient data for a context, smoothing becomes less important. As an example we trained the LEX-LDA model on BNC verb-object co-occurrences and estimated posterior mean values for σ_c for all verbs occurring more than 100 times and taking at least 10 different object argument types. The verbs with highest and lowest values are listed in Table 3. Although almost anything can be *discussed* or *highlighted*,

for topic $z \in \{1 \ldots |Z|\}$ **do**
  (Draw a distribution over words)
  $\boldsymbol{\Phi}_z \sim Dirichlet(\boldsymbol{\beta})$
end for
for context $c \in \{1 \ldots |\mathcal{C}|\}$ **do**
  (Draw a distribution over topics)
  $\boldsymbol{\theta}_c \sim Dirichlet(\boldsymbol{\alpha})$
  (Draw a distribution over words)
  $\boldsymbol{\Xi}_c \sim Dirichlet(\boldsymbol{\kappa})$
  (Draw a lexicalization probability)
  $\sigma_c \sim Beta(\lambda_0, \lambda_1)$
  for observation $o_i \in O(c)$ **do**
    (Draw a lexicalization indicator)
    $s_i \sim Bernoulli(\sigma_c)$
    if $s_i = 0$ **then**
      (Draw a topic)
      $z_i \sim Multinomial(\boldsymbol{\theta}_c)$
      (Draw a word)
      $w_i \sim Multinomial(\boldsymbol{\Phi}_{z_i})$
    else
      (Draw a word)
      $w_i \sim Multinomial(\boldsymbol{\Xi}_c)$
    end if
  end for
end for

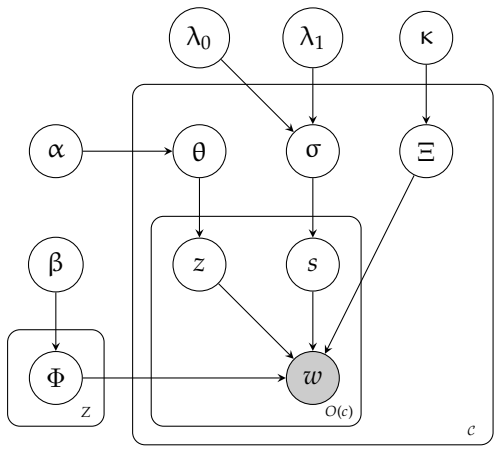


Figure 3
Generative story and plate diagram for LEX-LDA.

Table 3
BNC verbs with lowest and highest estimated lexicalization values σ_c for their object arguments, as well as the arguments with highest $P_{lex}(w|c)$ for high-lexicalization verbs.

Lowest σ_c		Highest σ_c		Top lexicalized arguments
discuss	1.2×10^{-4}	pose	0.872	problem, threat, question, challenge, risk
highlight	4.6×10^{-4}	wreak	0.864	havoc, vengeance, revenge, damage
consume	5.4×10^{-4}	adjourn	0.857	case, hearing, meeting, inquest, trial
emphasize	5.8×10^{-4}	reap	0.857	benefit, rewards, harvest, advantage
assert	6.5×10^{-4}	exert	0.851	influence, pressure, effect, control, force
contrast	6.5×10^{-4}	retrace	0.847	step, route, footstep, path, journey
obscure	6.8×10^{-4}	solve	0.847	problem, mystery, equation, crisis, case
document	6.8×10^{-4}	sip	0.839	coffee, tea, drink, wine, champagne
debate	6.9×10^{-4}	answer	0.826	question, call, phone, door, query
safeguard	8.0×10^{-4}	incur	0.823	cost, expense, loss, expenditure, liability

verbs such as *pose* and *wreak* have very lexicalized argument preferences. The semantic classes learned by LEX-LDA are broadly comparable to those learned by LDA, though it is less likely to mix classes on the basis of a single argument lexicalization; whereas the LDA class in row 5 of Table 1 is distracted by the high-frequency collocations *nature reserve* and *raise eyebrow*, LEX-LDA models trained on the same data can explain these through lexicalization effects and separate out body parts, conservation areas, and investments in different classes.

3.4 Parameter and Hyperparameter Learning

3.4.1 Learning Methods. A variety of methods are available for parameter learning in Bayesian models. The two standard approaches are variational inference, in which an approximation to the true distribution over parameters is estimated exactly, and sampling, in which convergence to the true posterior is guaranteed in theory but rarely verifiable in practice. In some cases the choice of approach is guided by the model, but often it is a matter of personal preference; for LDA, there is evidence that equivalent levels of performance can be achieved through variational learning and sampling given appropriate parameterization (Asuncion et al. 2009). In this article we use learning methods based on Gibbs sampling, following Griffiths and Steyvers (2004). The basic idea of Gibbs sampling is to iterate through the corpus one observation at a time, updating the latent variable value for each observation according to the conditional probability distribution determined by the current observed and latent variable values for all other observations. Because the likelihoods are multinomials with Dirichlet priors, we can integrate out their parameters using Equation (21).

For LDA, the conditional probability that the latent variable for the ith observation is assigned value z is computed as

$$P(z_i = z|\mathbf{z}^{-i}, c_i, w_i) \propto (f_{zc_i} + \alpha_z)\frac{f_{zw_i} + \beta}{f_z + |\mathcal{W}|\beta} \tag{30}$$

where $\mathbf{z}^{-i}$ is the set of current assignments for all observations other than the ith, f_z is the number of observations in that set assigned latent variable z, f_{zc_i} is the number of observations with context c_i assigned latent variable z, and f_{zw_i} is the number of observations with word w_i assigned latent variable z.

For ROOTH-LDA we make a similar calculation:

$$P(z_i = z|\mathbf{z}^{-i}, c_i, w_i) \propto (f_z + \alpha_z)\frac{f_{zw_i} + \beta}{f_z + |\mathcal{W}|\beta}\frac{f_{zc_i} + \beta}{f_z + |\mathcal{C}|\gamma} \tag{31}$$

For LEX-LDA the lexicalization variables s_i must also be sampled for each token. We "block" the sampling for z_i and s_i to improve convergence. The Gibbs sampling distribution is

$$P(s_i = 0, z_i = z|\mathbf{z}^{-i}, \mathbf{s}^{-i}, c_i, w_i) \propto (f_{c_i,s=0} + \lambda_0)\frac{f_{zc_i} + \alpha_z}{f_{c_i,s=0} + \sum_{z'} \alpha_{z'}}\frac{f_{zw_i} + \beta}{f_z + |\mathcal{W}|\beta} \tag{32}$$

$$P(s_i = 1, z_i = \emptyset|\mathbf{z}^{-i}, \mathbf{s}^{-i}, c_i, w_i) \propto (f_{c_i,s=1} + \lambda_1)\frac{f_{c_i w_i,s=1} + \kappa}{f_{c_i,s=1} + |\mathcal{W}|\kappa} \tag{33}$$

$$P(s_i = 0, z_i = \emptyset|\mathbf{z}^{-i}, \mathbf{s}^{-i}, c_i, w_i) = 0 \tag{34}$$

$$P(s_i = 1, z_i \neq \emptyset|\mathbf{z}^{-i}, \mathbf{s}^{-i}, c_i, w_i) = 0 \tag{35}$$

where $\emptyset$ indicates that no topic is assigned. The fact that topics are not assigned for all tokens means that LEX-LDA is less useful in situations that require representational power they afford—for example, the contextual similarity paradigm described in Section 3.5.

A naive implementation of the sampler will take time linear in the number of topics and the number of observations to complete one iteration. Yao, Mimno, and McCallum (2009) present a new sampling algorithm for LDA that yields a considerable speedup by reformulating Equation (30) to allow caching of intermediate values and an intelligent sorting of topics so that in many cases only a small number of topics need be iterated though before assigning a topic to an observation. In this article we use Yao, Mimno, & McCallum's algorithm for LDA, as well as a transformation of the ROOTH-LDA and LEX-LDA samplers that can be derived in an analogous fashion.

3.4.2 Inference. As noted previously, the Gibbs sampling procedure is guaranteed to converge to the true posterior after a finite number of iterations; however, this number is unknown and it is difficult to detect convergence. In practice, we run the sampler for a hopefully sufficient number of iterations and perform inference based on the final sampling state (assignments of all z and s variables) and/or a set of intermediate sampling states.

In the case of the LDA model, the selectional preference probability $P(w|c)$ is estimated using posterior mean estimates of θ_c and Φ_z:

$$P(w|c) = \sum_z P(z|c)P(w|z) \tag{36}$$

$$P(z|c) = \frac{f_{zc} + \alpha_z}{f_c + \sum_{z'} \alpha_{z'}} \tag{37}$$

$$P(w|z) = \frac{f_{zw} + \beta}{f_z + |\mathcal{W}|\beta} \tag{38}$$

For ROOTH-LDA, the joint probability $P(c, w)$ is given by

$$P(c, w) = \sum_z P(c|z)P(w|z)P(z) \tag{39}$$

$$P(z) = \frac{f_z + \alpha_z}{|O| + \sum_{z'} \alpha_{z'}} \tag{40}$$

$$P(w|z) = \frac{f_{zw} + \beta}{f_z + |\mathcal{W}|\beta} \tag{41}$$

$$P(c|z) = \frac{f_{zc} + \gamma}{f_z + |\mathcal{C}|\gamma} \tag{42}$$

For LEX-LDA, $P(w|c)$ is given by

$$P(w|c) = P(\sigma = 1|c)P_{lex}(w|c) + P(\sigma = 0|c)P_{class}(w|c) \tag{43}$$

$$P(\sigma = 1|c) = \frac{f_{c,s=1} + \lambda_1}{f_c + \lambda_0 + \lambda_1} \tag{44}$$

$$P(\sigma = 0|c) = 1 - P(\sigma = 1|c) \tag{45}$$

$$P_{lex}(w|c) = \frac{f_{wc,s=1} + \kappa}{f_{c,s=1} + |\mathcal{W}|\kappa} \tag{46}$$

$$P_{class}(w|c) = \sum_z P(z|c)P(w|z) \tag{47}$$

$$P(z|c) = \frac{f_{zc} + \alpha_z}{f_{c,s=0} + \sum_{z'} \alpha_{z'}} \tag{48}$$

$$P(w|z) = \frac{f_{zw} + \beta}{f_z + |\mathcal{W}|\beta} \tag{49}$$

Given a sequence or **chain** of sampling states $S_1, \ldots, S_n$, we can predict a value for $P(w|c)$ or $P(c, w)$ using these equations and the set of latent variable assignments at a single state S_i. As the sampler is initialized randomly and will take time to find a good area of the search space, it is standard to wait until a number of iterations have passed before using any samples for prediction. States $S_1, \ldots, S_b$ from this **burn-in** period are discarded.

For predictive stability it can be beneficial to average over predictions computed from more than one sampling state; for example, we can produce an averaged estimate of $P(w|c)$ from a set of states S:

$$P(w|c) = \frac{1}{|S|} \sum_{S_i \in S} P_{S_i}(w|c) \tag{50}$$

It is also possible to average over states drawn from multiple chains. However, averaging of any kind can only be performed on quantities whose interpretation does not depend on the sampling state itself. For example, we cannot average over estimates of $P(z_1|c)$ drawn from different samples as the topic called z_1 in one iteration is not identical to the topic called z_1 in another; even within the same chain, the meaning of a topic will often change gradually from state to state.

3.4.3 Choosing $|Z|$. In the "parametric" latent variable models used here the number of topics or semantic classes, $|Z|$, must be fixed in advance. This brings significant efficiency advantages but also the problem of choosing an appropriate value for $|Z|$. The more classes a model has, the greater its capacity to capture fine distinctions between entities. However, this finer granularity inevitably comes at a cost of reduced generalization. One approach is to choose a value that works well on training or development data before evaluating held-out test items. Results in lexical semantics are often reported over the entirety of a data set, meaning that if we wish to compare those results we cannot hold out any portion. If the method is relatively insensitive to the parameter it may be sufficient to choose a default value. Rooth et al. (1999) suggest cross-validating on the training data likelihood (and not on the ultimate evaluation measure). An alternative solution is to average the predictions of models trained with different choices of $|Z|$; this avoids the need to pick a default and can give better results than any one value as it integrates contributions at different levels of granularity. As mentioned in Section 3.4.2 we must take care when averaging predictions to compute with quantities that do not rely on topic identity—for example, estimates of $P(a|p)$ can safely be combined whereas estimates of $P(z_1|p)$ cannot.

3.4.4 Hyperparameter Estimation. Although the likelihood parameters can be integrated out, the parameters for the Dirichlet and Beta priors (often referred to as "hyperparameters") cannot and must be specified either manually or automatically. The value of these parameters affects the sparsity of the learned posterior distributions. Furthermore, the use of an asymmetric prior (where not all its parameters have equal value) implements an assumption that some observation values are more likely than others before any observations have been made. Wallach, Mimno, and McCallum (2009) demonstrate that the parameterization of the Dirichlet priors in an LDA model has a material effect on performance, recommending in conclusion a symmetric prior on the "emission" likelihood $P(w|z)$ and an asymmetric prior on the document topic likelihoods $P(z|d)$. In this article we follow these recommendations and, like Wallach, Mimno, and McCallum, we optimize the relevant hyperparameters using a fixed point iteration to maximize the log evidence (Minka 2003; Wallach 2008).

3.5 Measuring Similarity in Context with Latent-Variable Models

The representation induced by latent variable selectional preference models also allows us to capture the disambiguatory effect of context. Given an observation of a word in a context, we can infer the most probable semantic classes to appear in that context and we can also infer the probability that a class generated the observed word. We can also estimate the probability that the semantic classes suggested by the observation would have licensed an alternative word. Taken together, these can be used to estimate in-context semantic similarity. The fundamental intuitions are similar to those behind the vector-space models in Section 2.3.2, but once again we are viewing them from the perspective of probabilistic modeling.

The basic idea is that we identify the similarity between an observed term w_o and an alternative term w_s in context C with the similarity between the probability distribution over latent variables associated with w_o and C and the probability distribution over latent variables associated with w_s:

$$sim(w_o, w_s|C) = sim(P(z|w_o, C), P(z|w_s)) \tag{51}$$

This assumes that we can associate a distribution over the same set of latent variables with each context item $c \in C$. As noted in Section 2.3.2, previous research has found that conditioning the representation of both the observed term and the candidate substitute on the context gives worse performance than conditioning the observed term alone; we also found a similar effect. Dinu and Lapata (2010) present a specific version of this framework, using a window-based definition of context and the assumption that the similarity given a set of contexts is the product of the similarity value for each context:

$$sim_{DL10}(w_o, w_s|C) = \prod_{c \in C} sim(P(z|w_o, c), P(z|w_s)) \tag{52}$$

In this article we generalize to syntactic as well as window-based contexts and also derive a well-motivated approach to incorporating multiple contexts inside the probability model; in Section 4.5 we show that both innovations contribute to improved performance on a lexical substitution data set.

The distributions we use for prediction are as follows. Given an LDA latent variable preference model that generates words given a context, it is straightforward to calculate the distribution over latent variables conditioned on an observed context–word pair:

$$P_{C \to T}(z|w_o, c) = \frac{P(w_o|z)P(z|c)}{\sum_{z'} P(w_o|z')P(z'|c)} \tag{53}$$

Given a set of multiple contexts C, each of which has an opinion about the distribution over latent variables, this becomes

$$P(z|w_o, C) = \frac{P(w_o|z)P(z|C)}{\sum_{z'} P(w_o|z')P(z'|C)} \tag{54}$$

$$P(z|C) = \frac{\prod_{c \in C} P(z|c)}{\sum_{z'} \prod_{c \in C} P(z'|c)} \tag{55}$$

The uncontextualized distribution $P(z|w_s)$ is not given directly by the LDA model. It can be estimated from relative frequencies in the Gibbs sampling state; we use an unsmoothed estimate.[8] We denote this model $C \to T$ to note that the target word is generated given the context.

Where the context–word relationship is asymmetric (as in the case of syntactic dependency contexts), we can alternatively learn a model that generates contexts given a target word; we denote this model $T \to C$:

$$P_{T \to C}(z|w_o, c) = \frac{P(z|w_o)P(c|z)}{\sum_{z'} P(z'|w_o)P(c|z')} \tag{56}$$

Again, we can generalize to non-singleton context sets:

$$P(z|w_o, C) = \frac{P(z|w_o)P(C|z)}{\sum_{z'} P(z'|w_o)P(C|z')} \tag{57}$$

8 In the notation of Section 3.4, this estimate is given by $\frac{f_{zw_s}}{f_{w_s}}$.

where

$$P(C|z) = \prod_{c \in C} P(c|z) \tag{58}$$

Equation (57) has the form of a "product of experts" model (Hinton 2002), though unlike many applications of such models we train the experts independently and thus avoid additional complexity in the learning process. The uncontextualized distribution $P(z|w_s)$ is an explicit component of the $T \rightarrow C$ model.

An analogous definition of similarity can be derived for the ROOTH-LDA model. Here there is no asymmetry as the context and target are generated jointly. The distribution over topics for a context c and target word w_o is given by

$$P_{\text{ROOTH-LDA}}(z|w_o, c) = \frac{P(w_o, c|z)P(z)}{\sum_{z'} P(w_o, c|z')P(z')} \tag{59}$$

while calculating the uncontextualized distribution $P(z|w_s)$ requires summing over the set of possible contexts C′:

$$P_{\text{ROOTH-LDA}}(z|w_s) = \frac{P(z) \sum_{c' \in C'} P(w_s, c|z)}{\sum_{z'} P(z') \sum_{c' \in C'} P(w_s, c|z')} \tag{60}$$

Because the interaction classes learned by ROOTH-LDA are specific to a relation type, this model is less applicable than LDA to problems that involve a rich context set C.

Finally, we must choose a measure of similarity between probability distributions. The information theory literature has provided many such measures; in this article we use the Bhattacharyya coefficient (Bhattacharyya 1943):

$$sim_{bhatt}(P_x(\mathbf{z}), P_y(\mathbf{z})) = \sum_z \sqrt{P_x(z)P_y(z)} \tag{61}$$

One could alternatively use similarities derived from probabilistic divergences such as the Jensen–Shannon Divergence or the L_1 distance (Lee 1999; Ó Séaghdha and Copestake 2008).

3.6 Related Work

As related earlier, non-Bayesian mixture or latent-variable approaches to co-occurrence modeling were proposed by Pereira, Tishby, and Lee (1993) and Rooth et al. (1999). Blitzer, Globerson, and Pereira (2005) describe a co-occurrence model based on a different kind of distributed latent-variable architecture similar to that used in the literature on neural language models. Brody and Lapata (2009) use the clustering effects of LDA to perform word sense induction. Vlachos, Korhonen, and Ghahramani (2009) use non-parametric Bayesian methods to cluster verbs according to their co-occurrences with subcategorization frames. Reisinger and Mooney (2010, 2011) have also investigated Bayesian methods for lexical semantics in a spirit similar to that adopted here. Reisinger and Mooney (2010) describe a "tiered clustering" model that, like LEX-LDA, mixes a cluster-based preference model with a predicate-specific distribution over

words; however, their model does not encourage sharing of classes between different predicates. Reisinger and Mooney (2011) propose a very interesting variant of the latent-variable approach in which different kinds of contextual behavior can be explained by different "views," each of which has its own distribution over latent variables; this model can give more interpretable classes than LDA for higher settings of $|Z|$.

Some extensions of the LDA topic model incorporate local as well as document context to explain lexical choice. Griffiths et al. (2004) combine LDA and a hidden Markov model (HMM) in a single model structure, allowing each word to be drawn from either the document's topic distribution or a latent HMM state conditioned on the preceding word's state; Moon, Erk, and Baldridge (2010) show that combining HMM and LDA components can improve unsupervised part-of-speech induction. Wallach (2006) also seeks to capture the influence of the preceding word, while at the same time generating every word from inside the LDA model; this is achieved by conditioning the distribution over words on the preceding word type as well as on the chosen topic. Boyd-Graber and Blei (2008) propose a "syntactic topic model" that makes topic selection conditional on both the document's topic distribution and on the topic of the word's parent in a dependency tree. Although these models do represent a form of local context, they either use a very restrictive one-word window or a notion of syntax that ignores lexical or dependency-label effects; for example, knowing that the head of a noun is a verb is far less informative than knowing that the noun is the direct object of *eat*.

More generally, there is a connection between the models developed here and latent-variable models used for parsing (e.g., Petrov et al. 2006). In such models each latent state corresponds to a "splitting" of a part-of-speech label so as to produce a finer-grained grammar and tease out intricacies of word–rule "co-occurrence." Finkel, Grenager, and Manning (2007) and Liang et al. (2007) propose a non-parametric Bayesian treatment of state splitting. This is very similar to the motivation behind an LDA-style selectional preference model. One difference is that the parsing model must explain the parse tree structure as well as the choice of lexical items; another is that in the selectional preference models described here each head–dependent relation is treated as an independent observation (though this could be changed). These differences allow our selectional preference models to be trained efficiently on large corpora and, by focusing on lexical choice rather than syntax, to home in on purely semantic information. Titov and Klementiev (2011) extend the idea of latent-variable distributional modeling to do "unsupervised semantic parsing" and reason about classes of semantically similar lexicalized syntactic fragments.

4. Experiments

4.1 Training Corpora

In our experiments we use two training corpora:

BNC the written component of the British National Corpus,[9] comprising around 90 million words. The corpus was tagged for part of speech, lemmatized, and parsed with the RASP toolkit (Briscoe, Carroll, and Watson 2006).

9 http://www.natcorp.ox.ac.uk/.

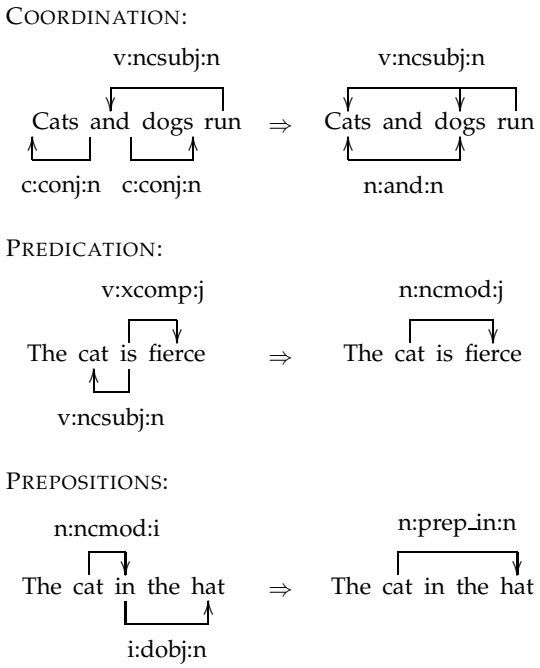


Figure 4
Dependency graph preprocessing.

WIKI a Wikipedia dump of over 45 million sentences (almost 1 billion words) tagged, lemmatized, and parsed with the C+C toolkit[10] and the fast CCG parser described by Clark et al. (2009).

Although two different parsers were used, they both have the ability to output grammatical relations in the RASP format and hence they are interoperable for our purposes as downstream users. This allows us to construct a combined corpus by simply concatenating the **BNC** and **WIKI** corpora.

In order to train our selectional preference models, we extracted word–context observations from the parsed corpora. Prior to extraction, the dependency graph for each sentence was transformed using the preprocessing steps illustrated in Figure 4. We then filtered for semantically discriminative information by ignoring all words with part of speech other than common noun, verb, adjective, and adverb. We also ignored instances of the verbs *be* and *have* and discarded all words containing non-alphabetic characters and all words with fewer than three characters.[11]

As mentioned in Section 2.1, the distributional semantics framework admits flexibility in how the practitioner defines the context of a word w. We investigate two possibilities in this article:

Syn The context of w is determined by the syntactic relations r and words w' incident to it in the sentence's parse tree, as illustrated in Section 3.1.

10 `http://svn.ask.it.usyd.edu.au/trac/candc`.
11 An exception was made for the word *PC* as it appears in the Keller and Lapata (2003) data set used for evaluation.

Win5 The context of w is determined by the words appearing within a window of five words on either side of it. There are no relation labels, so there is essentially just one relation r to consider.

Training topic models on a data set with very large "documents" leads to tractability issues. The window-based approach is particularly susceptible to an explosion in the number of extracted contexts, as each token in the data can contribute $2 \times W$ word–context observations, where W is the window size. We reduced the data by applying a simple downsampling technique to the training corpora. For the **WIKI/Syn** corpus, all word–context counts were divided by 5 and rounded to the nearest integer. For the **WIKI/Win5** corpus we divided all counts by 70; this number was suggested by Dinu and Lapata (2010), who used the same ratio for downsampling the similarly sized English Gigaword Corpus. Being an order of magnitude smaller, the BNC required less pruning; we divided all counts in the **BNC/Win5** by 5 and left the **BNC/Syn** corpus unaltered. Type/token statistics for the resulting sets of observations are given in Table 4.

4.2 Evaluating Selectional Preference Models

Various approaches have been suggested in the literature for evaluating selectional preference models. One popular method is "pseudo-disambiguation," in which a system must distinguish between actually occurring and randomly generated predicate–argument combinations (Pereira, Tishby, and Lee 1993; Chambers and Jurafsky 2010). In a similar vein, probabilistic topic models are often evaluated by measuring the probability they assign to held-out data; held-out likelihood has also been used for evaluation in a task involving selectional preferences (Schulte im Walde et al. 2008). These two approaches take a "language modeling" approach in which model quality is identified with the ability to predict the distribution of co-occurrences in unseen text. Although this metric should certainly correlate with the semantic quality of the model, it may also be affected by frequency and other idiosyncratic aspects of language use unless tightly controlled. In the context of document topic modeling, Chang et al. (2009) find that a model can have better predictive performance on held-out data while inducing topics that human subjects judge to be less semantically coherent.

In this article we choose to evaluate models by comparing system predictions with semantic judgments elicited from human subjects. These judgments take various forms. In Section 4.3 we use judgments of how plausible it is that a given predicate takes a given word as its argument. In Section 4.4 we use judgments of similarity

Table 4
Type and token counts for the **BNC** and **BNC+WIKI** corpora.

	BNC			**BNC+WIKI**		
	Tokens	Types	Contexts	Tokens	Types	Contexts
Nouns	18,723,082	122,999	316,237	54,145,216	106,448	514,257
Verbs	7,893,462	18,494	57,528	20,082,658	16,673	82,580
Adjectives	4,385,788	73,684	37,163	11,536,424	88,488	57,531
Adverbs	1,976,837	7,124	14,867	3,017,936	4,056	18,510
Window5	28,329,238	88,265	102,792	42,828,094	139,640	143,443

between pairs of predicate–argument combinations. In Section 4.5 we use judgments of substitutability for a target word as disambiguated by its sentential context. Taken together, these different experimental designs provide a multifaceted analysis of model quality.

4.3 Predicate–Argument Plausibility

4.3.1 Data. For the plausibility-based evaluation we use a data set of human judgments collected by Keller and Lapata (2003). This comprises data for three grammatical relations: verb–object, adjective–noun, and noun–noun modification. For each relation, 30 predicates were selected; each predicate was paired with three noun arguments from different predicate–argument frequency bands in the BNC as well as three noun arguments that were not observed for that predicate in the BNC. In this way two subsets (*Seen* and *Unseen*) of 90 items each were assembled for each predicate. Human plausibility judgments were elicited from a large number of subjects; these numerical judgments were then normalized, log-transformed, and averaged in a Magnitude Estimation procedure.

Predicate	Seen		Unseen	
dredge	*channel*	0.1875	*legend*	–0.3221
dredge	*canal*	0.2388	*sheet*	–0.2486
dredge	*rubbish*	–0.1999	*survivor*	–0.2077

Following Keller and Lapata (2003), we evaluate our models by measuring the correlation between system predictions and the human judgments. Keller and Lapata use Pearson's correlation coefficient r; we additionally use Spearman's rank correlation coefficient ρ for a non-parametric evaluation. Each system prediction is log-transformed before calculating the correlation to improve the linear fit to the gold standard.

4.3.2 Methods. We evaluate the LDA, ROOTH-LDA, and LEX-LDA latent-variable preference models, trained on predicate–argument pairs (c, w) extracted from the BNC. We use a default setting $|Z| = 100$ for the number of classes; in our experiments we have observed that our Bayesian models are relatively robust to the choice of $|Z|$. We average predictions of the joint probability $P(c, w)$ over three independent samples, each of which is obtained by sampling $P(c, w)$ every 50 iterations after a burn-in period of 200 iterations. ROOTH-LDA gives joint probabilities by definition (25), but LDA and LEX-LDA are defined in terms of conditional probabilities (24). There are two options for training these models:

$P \rightarrow A$**:** Model the distribution $P(w|c)$ over arguments for each predicate.
$A \rightarrow P$**:** Model the distribution $P(c|w)$ over predicates for each argument.

As the descriptions suggest, the definition of "predicate" and "argument" is arbitrary; it is equally valid to talk of the selectional preference of a noun for verbs taking it as a direct object as it is to talk of the preference of a verb for nouns taking it as a direct object. We expect both configurations to perform comparably on average, though there

may be linguistic or conceptual reasons why one configuration is better than the other for specific classes of co-occurrence.

To convert conditional probabilities to joint probabilities we multiply by a relative-frequency (MLE) estimate of the probability of the conditioning term:

$$P_{P \to A} = P(w|c)P(c) \tag{62}$$

$$P_{A \to P} = P(c|w)P(w) \tag{63}$$

As well as evaluating $P \to A$ and $A \to P$ implementations of LDA and LEX-LDA, we can evaluate a combined model $P \leftrightarrow A$ that simply averages the two sets of predictions; this removes the arbitrariness involved in choosing one direction or the other.

For comparison, we report the performance figures given by Keller and Lapata for their search-engine method using AltaVista and Google[12] as well as a number of alternative methods that we have reimplemented and trained on identical data:

BNC (MLE) A maximum-likelihood estimate proportional to the co-occurrence frequency $f(c, w)$ in the parsed BNC.

BNC (KN) BNC relative frequencies smoothed with modified Kneser-Ney (Chen and Goodman 1999).

Resnik The WordNet-based association strength of Resnik (1993). We used WordNet version 2.1 as the method requires multiple roots in the hierarchy for good performance.

Clark/Weir The WordNet-based method of Clark and Weir (2002), using WordNet 3.0. This method requires that a significance threshold α and significance test be chosen; we investigated a variety of settings and report performance for $\alpha = 0.9$ and Pearson's χ^2 test, as this combination consistently gave the best results.

Rooth-EM Rooth et al. (1999)'s latent-variable model without priors, trained with EM. As for the Bayesian models, we average the predictions over three iterations. This method is very sensitive to the number of classes; as proposed by Rooth et al., we choose the number of classes from the range (20, 25, . . . , 50) through 5-fold cross-validation on a held-out log-likelihood measure.

EPP The vector-space method of Erk, Padó, and Padó (2010), as described in Section 2.2.3. We used the cosine similarity measure for smoothing as it performed well in Erk, Padó, & Padó's experiments.

Disc A discriminative model inspired by Bergsma, Lin, and Goebel (2008) (see Section 2.2.4). In order to get true probabilistic predictions, we used a logistic regression classifier with L_1 regularization rather than a Support Vector Machine.[13] We train one classifier per predicate in the Keller and Lapata data set. Following Bergsma, Lin, and Goebel, we generate pseudonegative instances for each predicate by sampling noun arguments that either do not co-occur with it or have a negative PMI association. Again following Bergsma, Lin, and Goebel, we use a ratio of two pseudonegative instances for each positive instance and require pseudonegative arguments to be in the same frequency quintile as the matched

12 Keller and Lapata only report Pearson's r correlations; as we do not have their per-item predictions we cannot calculate Spearman's ρ correlations or statistical significance scores.

13 We used the logistic regression implementation provided by LIBLINEAR (Fan et al. 2008), available at `http://www.csie.ntu.edu.tw/~cjlin/liblinear`.

observed argument. The features used for each data instance, corresponding to an argument, are: the conditional probability of the argument co-occurring with each predicate in the training data; and string-based features capturing the length and initial and final character n-grams of the argument word.[14] We also investigate whether our LDA model can be used to provide additional features for the discriminative model, by giving the index of the most probable class $\max_z P(z|c, w)$; results for this system are labeled **Disc+LDA**.

In order to test statistical significance of performance differences we use a test for correlated correlation coefficients proposed by Meng, Rosenthal, and Rubin (1992). This is more appropriate than a standard test for independent correlation coefficients as it takes into account the strength of correlation between two sets of system outputs as well as each output's correlation with the gold standard. Essentially, if the two sets of system outputs are correlated there is less chance that their difference will be deemed significant. As we have no a priori reason to believe that one model will perform better than another, all tests are two-tailed.

4.3.3 Results. Results on the Keller and Lapata (2003) plausibility data set are presented in Table 5.[15] For common combinations (the Seen data) it is clear that relative corpus frequency is a reliable indicator of plausibility, especially when Web-scale resources are available. The BNC MLE estimate outperforms the best selectional preference model on three out of six Seen evaluations, and the AltaVista and Google estimates from Keller and Lapata (2003) outperforms the best selectional preference model on every applicable Seen evaluation. For the rarer Unseen combinations, however, MLE estimates are not sufficient and the latent-variable selectional preference models frequently outperform even the Web-based predictions. The results for BNC(KN) improve on the MLE estimates for the Unseen data but do not match the models that have a semantic component.

It is clear from Table 5 that the new Bayesian latent-variable models outperform the previously proposed selectional preference models under almost every evaluation. Among the latent-variable models there is no one clear winner, and small differences in performance are as likely to arise through random sampling variation as through qualitative differences between models. That said, ROOTH-LDA and LEX-LDA do score higher than LDA in a majority of cases. As expected, the bidirectional $P \leftrightarrow A$ models tend to perform at around the midpoint of the $P \rightarrow A$ and $A \rightarrow P$ models, though they can also exceed both; this suggests that they are a good choice when there is no intuitive reason to choose one direction over the other.

Table 6 aggregates comparisons for all combinations of the six data sets and two evaluation measures. As before, all the Bayesian latent-variable models achieve a roughly similar level of performance, consistently outperforming the models selected from the literature and frequently reaching statistical significance ($p < 0.05$). These results confirm that LDA-style models can be considered the current state of the art for selectional preference modeling.

14 Bergsma, Lin, and Goebel (2008) also use features extracted from gazetteers. However, they observe that additional features only give a small improvement over co-occurrence features alone. We do not use such features here but hypothesize that the improvement would be even smaller in our experiments as the data do not contain proper nouns.

15 Results for $\text{LDA}_{P \rightarrow A}$ and ROOTH-LDA were previously published in Ó Séaghdha (2010).

Table 5
Results (Pearson r and Spearman ρ correlations) on Keller and Lapata's (2003) plausibility data. Asterisks denote performance figures that are taken from the source paper; all other figures are drawn from our own (re)implementation trained on identical data.

	Verb–object				Noun–noun				Adjective–noun			
	Seen		Unseen		Seen		Unseen		Seen		Unseen	
	r	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ
AltaVista*	**.641**	–	.551	–	**.700**	–	.578	–	**.650**	–	.480	–
Google*	.624	–	.520	–	.692	–	.595	–	.641	–	.473	–
BNC (MLE)	.620	**.614**	.196	.222	.544	.604	.114	.125	.543	.622	.135	.102
BNC (KN)	.615	**.614**	.327	.350	.543	.594	.485	.523	.510	.619	.179	.173
Resnik	.384	.473	.469	.470	.242	.187	.152	.037	.309	.388	.311	.280
Clark/Weir	.489	.546	.312	.365	.441	.521	.543	.576	.440	.476	.271	.242
ROOTH-EM	.455	.487	.479	.520	.503	.491	.586	.625	.514	.463	.395	.355
EPP	.541	.562	.403	.436	.382	.465	.377	.398	.401	.400	.260	.195
Disc	.318	.318	.376	.354	.331	.294	.258	.250	.188	.274	.303	.327
Disc+LDA	.328	.338	.473	.476	.308	.285	.266	.292	.228	.308	.333	.368
$LDA_{P\rightarrow A}$	.504	.541	.558	.603	.615	.641	.636	.666	.594	.558	.468	.459
$LDA_{A\rightarrow P}$	.514	.555	.448	.469	.623	**.652**	.648	.688	.547	.583	.465	.458
$LDA_{P\leftrightarrow A}$	.513	.546	.530	.542	.619	.645	.653	.697	.593	.570	.467	.445
ROOTH-LDA	.520	.548	.564	.605	.607	.622	**.691**	**.722**	.575	.599	**.501**	.469
LEX-$LDA_{P\rightarrow A}$	.570	.600	**.601**	**.662**	.511	.537	.677	.706	.600	.627	.465	.451
LEX-$LDA_{A\rightarrow P}$	.568	.572	.523	.542	.532	.568	.659	.703	.545	.623	.513	.477
LEX-$LDA_{P\leftrightarrow A}$	.575	.589	.560	.599	.553	.563	.669	.698	.572	**.629**	**.517**	**.497**
Human*	.604	–	.640	–	.641	–	.570	–	.630	–	.550	–

Table 6
Aggregate comparisons for the Keller and Lapata (2003) plausibility data set between latent-variable models (rows) and previously proposed selectional preference models (columns). Cell entries give the number of evaluations (out of 12) in which the latent-variable model outperformed the alternative method and the number in which the improvement was statistically significant.

	Resnik	Clark/Weir	ROOTH-EM	EPP	Disc
$LDA_{P\rightarrow A}$	12/5	11/8	12/6	10/5	12/4
$LDA_{A\rightarrow P}$	10/4	12/8	10/5	10/3	12/4
$LDA_{P\leftrightarrow A}$	12/4	11/9	12/6	10/5	12/5
ROOTH-LDA	12/6	12/7	12/7	10/5	12/5
LEX-$LDA_{P\rightarrow A}$	12/5	11/8	12/6	12/5	12/5
LEX-$LDA_{A\rightarrow P}$	10/4	12/8	10/5	12/5	12/6
LEX-$LDA_{P\leftrightarrow A}$	12/4	11/9	12/6	12/5	12/5

One way of performing error analysis for a given result is to decompose the correlation coefficient into a sum of per-item "pseudo-coefficients." For Pearson's r, the contribution for the ith item is

$$r_i = \frac{(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_j (x_j - \mu_x)^2}\sqrt{\sum_j (y_j - \mu_y)^2}} \tag{64}$$

Spearman's ρ is equivalent to the r correlation between ranks and so a similar quantity can be computed. Table 7 illustrates the items with highest and lowest contributions for one evaluation (Spearman's ρ on the Keller and Lapata Unseen data set). We have attempted to identify general factors that predict the difficulty of an item by measuring rank correlation between the per-item pseudo-coefficients and various corpus statistics. However, it has proven difficult to isolate reliable patterns. One finding is that arguments with high corpus frequency tend to incur larger errors for the $P \rightarrow A$ latent-variable models and ROOTH-LDA, whereas predicates with high corpus frequency tend to incur smaller errors; with the $A \rightarrow P$ the effect is lessened but not reversed, suggesting that part of the effect may be inherent in the data set rather than in the prediction model.

4.4 Predicate–Argument Similarity

4.4.1 Data. Mitchell and Lapata (2008, 2010) collected human judgments of similarity between pairs of predicates and arguments corresponding to minimal sentences. Mitchell and Lapata's explicit aim was to facilitate evaluation of general semantic compositionality models but their data sets are also suitable for evaluating predicate–argument representations.

Mitchell and Lapata (2008) used the BNC to extract 4 attested subject nouns for each of 15 verbs, yielding 60 reference combinations. Each verb–noun tuple was matched with two verbs that are synonyms of the reference verb in some contexts but not in

Table 7
Most- and least-accurately predicted items for the $\text{LDA}_{P \rightarrow A}$ models using per-item Spearman's ρ pseudo-coefficients on the unseen data set, with gold and predicted rank values.

Item	r_i	Gold	Pred	Item	r_i	Gold	Pred
influence worker	0.030	3	2	spend life	–0.012	63	1
originate miner	0.029	89	86	rank pc	–0.012	21	75
undergo container	0.027	90	83	deduct stage	–0.011	79	25
litter surface	0.027	7	3	sponsor embassy	–0.010	23	73
injure pilot	0.026	2	9	spend error	–0.007	80	30
desk tomato	0.028	87	87	guitar conviction	–0.012	82	25
pupil morale	0.026	3	9	towel fee	–0.011	11	65
landlord committee	0.025	12	1	workshop victim	–0.007	18	60
restoration specialist	0.025	1	12	opera recommendation	–0.006	6	54
cable manager	0.024	7	8	valuation afternoon	–0.005	70	32
superb character	0.032	2	1	tremendous newspaper	–0.014	13	72
scientific document	0.032	1	2	continuous clinic	–0.012	75	21
valid silk	0.031	89	89	lazy promoter	–0.012	24	79
naughty protocol	0.026	84	87	unfair coalition	–0.012	20	73
exciting can	0.026	87	84	lazy shadow	–0.010	74	24

Table 8
Sample items from the Mitchell and Lapata (2008) data set.

shoulder slump *shoulder slouch*	6, 7, 5, 5, 6, 5, 5, 7, 5, 5, 7, 5, 6, 6, 5, 6, 6, 6, 7, 5, 7, 6, 6, 5, 5, 5, 5, 6, 6, 7, 7, 7, 7, 7
shoulder slump *shoulder decline*	2, 5, 4, 4, 3, 3, 2, 3, 2, 1, 3, 3, 6, 5, 3, 2, 1, 1, 1, 7, 4, 4, 6, 3, 5, 6

Table 9
Sample items from the Mitchell and Lapata (2010) data set.

stress importance *emphasize need*	6, 7, 7, 5, 5, 7, 7, 7, 6, 5, 6, 7, 3, 7, 7, 6, 7, 7
ask man *stretch arm*	3, 1, 4, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 1, 1, 2, 1
football club *league match*	7, 6, 7, 6, 6, 5, 5, 3, 6, 6, 4, 5, 4, 6, 2, 7, 5, 5
education course *training program*	7, 7, 5, 5, 7, 5, 5, 7, 7, 4, 6, 2, 5, 6, 6, 7, 7, 4

others. In this way, Mitchell and Lapata created a data set of 120 pairs of predicate–argument combinations. Similarity judgments were obtained from human subjects for each pair on a Likert scale of 1–7. Examples of the resulting data items are given in Table 8. Mitchell and Lapata use six subjects' ratings as a development data set for setting model parameters and the remaining 54 subjects' ratings for testing. In this article we use the same split.

Mitchell and Lapata (2010) adopt a similar approach to data collection with the difference that instead of keeping arguments constant across combinations in a pair, both predicates and arguments vary across comparand combinations. They also consider a range of grammatical relations: verb–object, adjective–noun, and noun–noun modification. Human subjects rated similarity between predicate–argument combinations on a 1–7 scale as before; examples are given in Table 9. Inspection of the data suggests that the subjects' annotation may conflate semantic similarity and relatedness; for example, *football club* and *league match* are often given a high similarity score. Mitchell and Lapata again split the data into development and testing sections, the former comprising 54 subjects' ratings and the latter comprising 108 subjects' ratings.

Turney (2012) reports, on the basis of personal communication, that Mitchell and Lapata (2010) used an involved evaluation procedure that is not described in their original paper; for each grammatical relation, the annotators are partitioned in three groups and the Spearman's ρ correlation computed for each group is combined by averaging.[16] The analogous approach for the Mitchell and Lapata (2008) data set calculates a single ρ value by pairing of each annotator-item score with the system prediction for the appropriate item. Let $\mathbf{s}$ be the sequence of system predictions for $|I|$ items and $\mathbf{y_a}$

16 We do not compare against the system of Turney (2012) as Turney uses a different experimental design based on partitioning by phrases rather than annotators.

be the scores assigned by annotator $a \in A$ to those $|I|$ items. Then the "concatenated" correlation ρ_{cat} is calculated as follows:[17]

$$\mathbf{y_{cat}} = (y_{a_1 i_1}, y_{a_1 i_2}, \ldots, y_{a_1 i_{|I|}}, y_{a_2 i_1} \ldots, y_{a_{|A|} i_{|I|}}) \tag{65}$$

$$\mathbf{s_{cat}} = (s_1, s_2, \ldots, s_{|I|}, s_1, \ldots, s_{|I|}) \tag{66}$$

$$\rho_{cat} = \rho(\mathbf{s_{cat}}, \mathbf{y_{cat}}) \tag{67}$$

The length of the $\mathbf{y_{cat}}$ and $\mathbf{s_{cat}}$ sequences is equal to the total number of annotator-item scores. For the Mitchell and Lapata (2010) data set, a ρ_{cat} value is calculated for each of the three annotator groups and these are then averaged. As Turney observes, this approach seems to have the effect of underestimating model quality relative to the inter-annotator agreement figure, which is calculated as average intersubject correlation. Therefore, in addition to Mitchell and Lapata's ρ_{cat} evaluation, we also perform an evaluation that computes the average correlation ρ_{ave} between the system output and each individual annotator:

$$\rho_{ave} = \frac{1}{|A|} \sum_{a \in A} \rho(\mathbf{s}, \mathbf{y_a}) \tag{68}$$

4.4.2 Models. For the Mitchell and Lapata (2008) data set we train the following models on the BNC corpus:

LDA An LDA selectional preference model of verb–subject co-occurrence with similarity computed as described in Section 3.5. Similarity predictions $sim(n, o|c)$ are averaged over five runs. We consider three models of context–target interaction, which in this case corresponds to verb–subject interaction:

LDA$_{C \to T}$ Target generation is conditioned on the context, as in equation (53).
LDA$_{T \to C}$ Context generation is conditioned on the target, as in equation (56).
LDA$_{C \leftrightarrow T}$ An average of the predictions made by LDA$_{C \to T}$ and LDA$_{T \to C}$.

As before, we consider a default setting of $|Z| = 100$. As well as presenting results for an average over all predictors we investigate whether the choice of predictors can be optimized by using the development data to select the best subset of predictors.

Mult Pointwise multiplication (6) using **Win5** co-occurrences.

We also compare against the best figures reported in previous studies; these also used the BNC for training and so should be directly comparable:

M+L08 The best-performing system of Mitchell and Lapata (2008), combining an additive and a multiplicative model and using window-based co-occurrences.

SVS The best-performing system of Erk and Padó (2008); the Structured Vector Space model (8), parameterized to use window-based co-occurrences and raising the expectation vector values (7) to the 20th power (this parameter was optimized on the development data).

17 In practice the sequence of items is not the same for every annotator and the sequence of predictions $\mathbf{s}$ must be changed accordingly.

Table 10
Results (ρ_{ave} averaged across annotators) for the Mitchell and Lapata (2008) similarity data set.

Model	No Optimization	Optimized on Dev
$\lvert Z\rvert = 100$		
$\text{LDA}_{C\rightarrow T}$	0.34	0.35
$\text{LDA}_{T\rightarrow C}$	**0.39**	**0.41**
$\text{LDA}_{C\leftrightarrow T}$	**0.39**	**0.41**
Mult	0.15	–
Human*	0.40	–

For the Mitchell and Lapata (2010) data set we train the following models, again on the BNC corpus:

ROOTH-LDA/Syn A ROOTH-LDA model trained on the appropriate set of syntactic co-occurrences (verb–object, noun–noun modification, or adjective–noun), with the topic distribution calculated as in Equation (59).

LDA/Win5 An LDA model trained on the **Win5** window-based co-occurrences. Because all observations are modeled using the same latent classes, the distributions $P(z|o, c)$ (Equation (53)) for each word in the pair can be combined by taking a normalized product.

Combined This model averages the similarity prediction of the **ROOTH-LDA/Syn** and **LDA/Win5** models.

Mult Pointwise multiplication (6) using **Win5** co-occurrences.

We report results for an average over all predictors as well as for the subset that performs best on the development data. We also list results that were reported by Mitchell and Lapata:

M+L10/Mult A multiplicative model (6) using a vector space based on window co-occurrences in the BNC.

M+L10/Best The best result for each grammatical relation from any of the semantic spaces and combination methods tested by Mitchell and Lapata. Some of these methods require parameters to be set through optimization on the development set.[18]

4.4.3 Results. Results for the Mitchell and Lapata (2008) data set are presented in Tables 10 and 11.[19] The LDA preference models clearly outperform the previous state of the art of $\rho_{cat} = 0.27$ (Erk and Padó 2008), with the best simple average of predictors scoring $\rho_{cat} = 0.38$, $\rho_{ave} = 0.41$, and the best optimized combination scoring $\rho_{cat} = 0.39$, $\rho_{ave} = 0.41$. This is comparable to the average level of agreement between human judges estimated by Mitchell and Lapata's to be $\rho_{ave} = 0.40$. Optimizing on the development data consistently gave better performance than averaging over all predictors, though in most cases the differences are small.

18 Ultimately, however, none of the combination methods needing optimization outperform the parameter-free methods in Mitchell and Lapata's results.

19 The results in Table 10 were previously published in Ó Séaghdha and Korhonen (2011).

Table 11
Results (ρ_{cat}) for the Mitchell and Lapata (2008) similarity data set.

Model	No Optimization	Optimized on Dev
$\lvert Z\rvert = 100$		
$\text{LDA}_{C\rightarrow T}$	0.28	0.32
$\text{LDA}_{T\rightarrow C}$	**0.38**	**0.39**
$\text{LDA}_{C\leftrightarrow T}$	0.33	0.38
Mult	0.13	–
SVS*	0.27	–
M+L08*	0.19	–
Human*	0.40	–

Table 12
Results (ρ_{ave} averaged across annotators) for the Mitchell and Lapata (2010) similarity data set.

Model	No Optimization			Optimized on Dev		
	V–Obj	N–N	Adj–N	V–Obj	N–N	Adj–N
LDA/Win5	0.41	**0.56**	0.46	0.42	**0.58**	0.49
ROOTH-LDA/Syn	0.42	0.46	0.51	0.42	0.47	0.52
Combined	**0.44**	**0.56**	**0.53**	**0.46**	**0.58**	**0.55**
Mult/Win5	0.34	0.33	0.34	–	–	–
Human*	0.55	0.49	0.52	–	–	–

Results for the Mitchell and Lapata (2010) data set are presented in Tables 12 and Table 13.[20] Again the latent-variable models perform well, comfortably outperforming the **Mult** baseline, and with just one exception the **Combined** models surpass Mitchell and Lapata's reported results. Combining the syntactic co-occurrence model **ROOTH-LDA/Syn** and the window-based model **LDA/Win5** consistently gives the best performance, suggesting that the human ratings in this data set are sensitive to both strict similarity and a looser sense of relatedness. As Turney (2012) observes, the average-ρ_{cat}-per-group approach of Mitchell and Lapata leads to lower performance figures than averaging across annotators; with the latter approach (Table 12) the ρ_{ave} correlation values approach the level of human interannotator agreement for two of the three relations: noun–noun and adjective–noun modification.

4.5 Lexical Substitution

4.5.1 Data. The data set for the English Lexical Substitution Task (McCarthy and Navigli 2009) consists of 2,010 sentences sourced from Web pages. Each sentence features one of 205 distinct target words that may be nouns, verbs, adjectives, or adverbs. The sentences have been annotated by human judges to suggest semantically acceptable substitutes for their target words. Table 14 gives example sentences and annotations for the target verb *charge*. For the original shared task the data was divided into development and test

20 These results were not previously published.

Table 13
Results (ρ_{cat} averaged across groups) for the Mitchell and Lapata (2010) similarity data set.

Model	No Optimization			Optimized on Dev		
	V–Obj	N–N	Adj–N	V–Obj	N–N	Adj–N
LDA/Win5	0.37	**0.51**	0.42	0.37	**0.53**	0.44
ROOTH-LDA/Syn	0.37	0.42	0.45	0.37	0.43	0.47
Combined	0.39	**0.51**	**0.47**	**0.41**	**0.53**	**0.48**
Mult/Win5	0.31	0.30	0.30	–	–	–
M+L10/Mult*	0.37	0.49	0.46	–	–	–
M+L10/Best*	**0.40**	0.49	0.46	0.40	0.49	0.46
Human*	0.55	0.49	0.52	–	–	–

Table 14
Example sentences for the verb *charge* from the English Lexical Substitution Task.

Commission is the amount *charged* to execute a trade.
levy (2), *impose* (1), *take* (1), *demand* (1)

Annual fees are *charged* on a pro-rata basis to correspond with the standardized renewal date in December.
levy (2), *require* (1), *impose* (1), *demand* (1)

Meanwhile, George begins obsessive plans for his funeral…George, suspicious, *charges* to her room to confront them.
run (2), *rush* (2), *storm* (1), *dash* (1)

Realizing immediately that strangers have come, the animals *charge* them and the horses began to fight.
attack (5), *rush at* (1)

sections; in this article we follow subsequent work using parameter-free models and use the whole data set for testing.

The gold standard substitute annotations contain a number of multiword terms such as *rush at* and *generate electricity*. As it is impossible for a standard lexical distributional model to reason about such terms, we remove these substitutes from the gold standard.[21] We remove entirely the 17 sentences that have only multiword substitutes in the gold standard, as well as 7 sentences for which no gold annotations are provided. This leaves 1,986 sentences.

The original Lexical Substitution Task asked systems to propose substitutes from an unrestricted English vocabulary, though in practice all participants used lexical resources to constrain the set of substitutes considered. Most subsequent researchers using the Lexical Substitution data to evaluate models of contextual meaning have adopted a slightly different experimental design, in which systems are asked to rank only the list of attested substitutes for the target word in each sentence. For example,

21 Thater, Fürstenau, and Pinkal (2010, 2011) and Dinu and Lapata (2010) similarly remove multiword paraphrases (Georgiana Dinu, p.c.).

the list of substitute candidates for an instance of *charge* is the union of the substitute lists in the gold standard for every sentence containing *charge* as a target word: *levy*, *impose*, *take*, *demand*, *require*, *impose*, *run*, *rush*, *storm*, *dash*, *attack*,.... Evaluation of system predictions for a given sentence then involves comparing the ranking produced by the system with the implicit ranking produced by annotators, assuming that any candidates not attested for the sentence appear with frequency 0 at the bottom of the ranking. Dinu and Lapata (2010) use Kendall's τ_b, a standard rank correlation measure that is appropriate for data containing tied ranks. Thater, Fürstenau, and Pinkal (2010, 2011) use Generalized Average Precision (GAP), a precision-like measure originally proposed by Kishida (2005) for information retrieval:

$$GAP = \frac{\sum_{i=1}^{n} I(x_i) \frac{\sum_{k=1}^{i} x_k}{i}}{\sum_{j=1}^{R} I(y_j) \frac{\sum_{l=1}^{j} y_l}{j}} \tag{69}$$

where $x_1, \ldots, x_n$ are the ranked candidate scores provided by the system, $y_1, \ldots, y_R$ are the ranked scores in the gold standard and $I(x)$ is an indicator function with value 1 if $x > 0$ and 0 otherwise.

In this article we report both τ_b and GAP scores, calculated individually for each sentence and averaged. The open-vocabulary design of the original Lexical Substitution Task facilitated the use of other evaluation measures such as "precision out of ten": the proportion of the first 10 words in a system's ranked substitute list that are contained in the gold standard annotation for that sentence. This measure is not appropriate in the constrained-vocabulary scenario considered here; when there are fewer than 10 candidate substitutes for a target word, the precision will always be 1.

4.5.2 Models. We apply both window-based and syntactic models of similarity in context to the lexical substitution data set; we expect the latter to give more accurate predictions but to have incomplete coverage when a test sentence is not fully and correctly parsed or when the test lexical items were not seen in the appropriate contexts in training.[22] We therefore also average the predictions of the two model types in the hope of attaining superior performance with full coverage.

The models we train on the **BNC** and combined **BNC + WIKI** corpora are as follows:

Win5 An LDA model using 5-word-window contexts (so $|C| \leq 10$) and similarity $P(z|o, C)$ computed according to Equation (54).

$C \rightarrow T$ An LDA model using syntactic co-occurrences with similarity computed according to Equation (54).

$T \rightarrow C$ An LDA model using syntactic co-occurrences with similarity computed according to Equation (57).

$T \leftrightarrow C$ A model averaging the predictions of the $C \rightarrow T$ and $T \rightarrow C$ models.

Win5 + $C \rightarrow T$**, Win5 +** $T \rightarrow C$**, Win5 +** $T \leftrightarrow C$ A model averaging the predictions of **Win5** and the appropriate syntactic model.

TFP11 The vector-space model of Thater, Fürstenau, and Pinkal (2011). We report figures with and without backoff to lexical similarity between target and substitute words in the absence of a syntax-based prediction.

22 An LDA model cannot make an informative prediction of $P(z|o, C)$ if word o was never seen entering into at least one (unlexicalized) syntactic relation in C. Other syntactic models such as that of Thater, Fürstenau, and Pinkal (2011) face analogous restrictions.

Table 15
Results on the English Lexical Substitution Task data set; **boldface** denotes best performance at full coverage for each corpus.

	BNC			BNC + Wikipedia		
	GAP	τ_b	%Coverage	GAP	τ_b	%Coverage
Win5	44.5	0.17	100.0	44.8	0.17	100.0
$C \rightarrow T$	46.8	0.20	86.4	48.7	0.21	86.5
$T \rightarrow C$	47.2	0.21	86.4	49.3	0.22	86.5
$T \leftrightarrow C$	48.2	0.22	86.4	49.1	0.23	86.5
Win5 + $C \rightarrow T$	46.0	0.18	100.0	48.7	0.21	100.0
Win5 + $T \rightarrow C$	**48.6**	**0.21**	100.0	49.3	0.22	100.0
Win5 + $T \leftrightarrow C$	48.1	0.20	100.0	**49.5**	**0.23**	100.0
Baseline:						
No Context	43.8	0.16	100.0	43.7	0.15	100.0
No Similarity	39.7	0.14	100.0	40.3	0.14	100.0
TFP11:						
No Backoff	46.8	0.20	84.8	47.7	0.22	84.9
+Backoff	46.4	0.19	98.1	47.3	0.21	98.2

We also consider two baseline LDA models:

No Context A model that ranks substitutes n by computing the Bhattacharyya similarity between their topic distributions $P(z|n)$ and the target word topic distribution $P(z|o)$.

No Similarity A model that ranks substitutes n by their context-conditioned probability $P(n|C)$ only; this is essentially a language-modeling approach using syntactic "bigrams."

We report baseline results for the $T \leftrightarrow C$ syntactic model, but performance is similar with other co-occurrence types.

Predictions for the LDA models are averaged over five runs for each setting of $|Z|$ in the range $\{600, 800, 1000, 1200\}$. In order to test statistical significance of differences between models we use stratified shuffling (Yeh 2000).[23]

4.5.3 Results. Table 15 presents results on the Lexical Substitution Task data set.[24] As expected, the window-based LDA models attain good coverage but worse performance than the syntactic models. The combined model **Win5** + $T \leftrightarrow C$ trained on **BNC+WIKI** gives the best scores (GAP = 49.5, $\tau_b = 0.23$). Every combined model gives a statistically significant improvement ($p < 0.01$) over the corresponding window-based **Win5** model. Our **TFP11** reimplementation of Thater, Fürstenau, and Pinkal (2011) has slightly less than complete coverage, and performs worse than almost all combined LDA models. To compute statistical significance we only use the sentences for which **TFP11** made predictions; for both the **BNC** and **BNC+WIKI** corpora, the **Win5** + $T \leftrightarrow C$ model

23 We use the software package provided by Sebastian Padó at `http://www.nlpado.de/~sebastian/sigf.html`.

24 Results for the LDA models were reported in Ó Séaghdha and Korhonen (2011).

gives a statistically significant ($p < 0.05$) improvement over **TFP11** for both GAP and τ_b, while **Win5** + $T \rightarrow C$ gives a significant improvement for GAP and τ_b on the BNC training corpus. The no-context and no-similarity baselines are clearly worse than the full models; this difference is statistically significant ($p < 0.01$) for both training corpora and all models.

Table 16 breaks performance down across the four parts of speech used in the data set. Verbs appear to present the most difficult substitution questions and also demonstrate the greatest beneficial effect of adding syntactic disambiguation to the basic **Win5** model. The full **Win5** + $T \leftrightarrow C$ outperforms our reimplementation of Thater, Fürstenau, and Pinkal (2011) on all parts of speech for the GAP statistic and on verbs and adjectives for τ_b, scoring a tie on nouns and adverbs. Table 16 also lists results reported by Dinu and Lapata (2010) and Thater, Fürstenau, and Pinkal (2010, 2011) for their models trained on the English Gigaword Corpus. This corpus is of comparable size to the **BNC+WIKI** corpus, but we note that the results reported by Thater, Fürstenau, and Pinkal (2011) are better than those attained by our reimplementation, suggesting that uncontrolled factors such as choice of corpus, parser, or dependency representation may be responsible. Thater, Fürstenau, and Pinkal's (2011) results remain the best reported for this data set; our **Win5** + $T \leftrightarrow C$ results are better than Dinu and Lapata (2010) and Thater, Fürstenau, and Pinkal (2010) in this uncontrolled setting.

5. Conclusion

In this article we have shown that the probabilistic latent-variable framework provides a flexible and effective toolbox for distributional modeling of lexical meaning and gives state-of-the-art results on a number of semantic prediction tasks. One useful feature of this framework is that it induces a representation of semantic classes at the same time as it learns about selectional preference distributions. This can be viewed as a kind of coarse-grained sense induction or as a kind of concept induction. We have demonstrated that reasoning about these classes leads to an accurate method for calculating semantic similarity in context. By applying our models we attain state-of-the-art performance on a range of evaluations involving plausibility prediction, in-context similarity, and

Table 16
Performance by part of speech, with additional results from Thater, Fürstenau, and Pinkal (2010, 2011) and Dinu and Lapata (2010).

	Nouns		Verbs		Adjectives		Adverbs		Overall	
	GAP	τ_b	GAP	τ_b	GAP	τ_b	GAP	τ_b	GAP	τ_b
Win5/BNC+WIKI	46.0	0.16	38.9	0.14	44.0	0.18	54.0	0.22	44.8	0.17
Win5 + $T \leftrightarrow C$	50.7	0.22	45.1	0.20	48.8	0.24	55.9	0.24	49.5	0.23
TFP11 (+Backoff)	48.9	0.22	42.5	0.17	46.0	0.22	55.2	0.24	47.3	0.21
TFP10* (Model 1)	46.4	–	45.9	–	39.4	–	48.2	–	44.6	–
TFP10* (Model 2)	42.5	–	–	–	43.2	–	51.4	–	–	–
TFP11* (+Backoff)	52.9	–	48.8	–	51.1	–	55.3	–	51.7	–
DL10* (LDA)	–	0.16	–	0.14	–	0.17	–	0.21	–	0.16
DL10* (NMF)	–	0.15	–	0.14	–	0.16	–	0.26	–	0.16

lexical substitution. The three models we have investigated—LDA, ROOTH-LDA and LEX-LDA—all perform at a similar level for predicting plausibility, but in other cases the representation induced by one model may be more suitable than the others.

In future work, we anticipate that the same intuitions may lead to similarity accurate methods for other tasks where disambiguation is required; an obvious candidate would be traditional word sense disambiguation, perhaps in combination with the probabilistic WordNet-based preference models of Ó Séaghdha and Korhonen (2012). More generally, we expect that latent-variable models will prove useful in applications where other selectional preference models have been applied, for example, metaphor interpretation and semantic role labeling.

A second route for future work is to enrich the semantic representations that are learned by the model. As previously mentioned, probabilistic generative models are modular in the sense that they can be integrated in larger models. Bayesian methods for learning tree structures could be applied to learn taxonomies of semantic classes (Blei, Griffiths, and Jordan 2010; Blundell, Teh, and Heller 2010). Borrowing ideas from Bayesian hierarchical language modeling (Teh 2006), one could build a model of selectional preference and disambiguation in the context of arbitrarily long dependency *paths*, relaxing our current assumption that only the immediate neighbors of a target word affect its meaning. Our class-based preference model also suggests an approach to identifying regular polysemy alternation by finding class co-occurrences that repeat across words, offering a fully data-driven alternative to polysemy models based on WordNet (Boleda, Padó, and Utt 2012). In principle, any structure that can be reasoned about probabilistically, from syntax trees to coreference chains or semantic relations, can be coupled with a selectional preference model to incorporate disambiguation or lexical smoothing in a task-oriented architecture.

Acknowledgments

The work in this article was funded by the EPSRC (grant EP/G051070/1) and by the Royal Society. We are grateful to Frank Keller and Mirella Lapata for sharing their data set of plausibility judgments; to Georgiana Dinu, Karl Moritz Hermann, Jeff Mitchell, Sebastian Padó, and Andreas Vlachos for offering information and advice; and to the anonymous *Computational Linguistics* reviewers, whose suggestions have substantially improved the quality of this article.

References

Abney, Steven and Marc Light. 1999. Hiding a semantic hierarchy in a Markov model. In *Proceedings of the ACL-99 Workshop on Unsupervised Learning in NLP*, pages 1–8, College Park, MD.

Altmann, Gerry T. M. and Yuki Kamide. 1999. Incremental interpretation of verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Asuncion, Arthur, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI-09)*, pages 27–34, Montreal.

Bergsma, Shane, Dekang Lin, and Randy Goebel. 2008. Discriminative learning of selectional preferences from unlabeled text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-08)*, pages 59–68, Honolulu, HI.

Bhattacharyya, A. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–110.

Bicknell, Klinton, Jeffrey L. Elman, Mary Hare, Ken McRae, and Marta Kutas. 2010. Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, 63(4):489–505.

Blei, David M., Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation.

Journal of Machine Learning Research, 3:993–1,022.

Blitzer, John, Amir Globerson, and Fernando Pereira. 2005. Distributed latent variable models of lexical co-occurrences. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS-05)*, pages 25–32, Barbados.

Blundell, Charles, Yee Whye Teh, and Katherine A. Heller. 2010. Bayesian rose trees. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 65–72, Catalina Island, CA.

Boleda, Gemma, Sebastian Padó, and Jason Utt. 2012. Regular polysemy: A distributional model. In *Proceedings of *SEM-12*, pages 151–160, Montreal.

Boyd-Graber, Jordan and David M. Blei. 2008. Syntactic topic models. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS-08)*, pages 185–192, Vancouver.

Briscoe, Ted. 2006. An introduction to tag sequence grammars and the RASP system parser. Technical Report 662, Computer Laboratory, University of Cambridge.

Briscoe, Ted, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the ACL-06 Interactive Presentation Sessions*, pages 77–80, Sydney.

Brody, Samuel and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of EACL-09*, pages 103–111, Athens.

Chambers, Nathanael and Dan Jurafsky. 2010. Improving the use of pseudo-words for evaluating selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, pages 445–453, Uppsala.

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS-09)*, pages 288–296, Vancouver.

Chen, Stanley F. and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359–393.

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton de Gruyter, Berlin.

Ciaramita, Massimiliano and Mark Johnson. 2000. Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, pages 187–193, Saarbrücken.

Clark, Stephen, Ann Copestake, James R. Curran, Yue Zhang, Aurelie Herbelot, James Haggerty, Byung-Gyu Ahn, Curt Van Wyk, Jessika Roesner, Jonathan Kummerfeld, and Tim Dawborn. 2009. Large-scale syntactic processing: Parsing the Web. Technical report, Final Report of the 2009 JHU CLSP Workshop.

Clark, Stephen and David Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206.

Cordier, Brigitte. 1965. Factor-analysis of correspondences. In *Proceedings of the 1965 International Conference on Computational Linguistics (COLING-65)*, New York, NY.

Curran, James. 2003. *From Distributional to Semantic Similarity*. Ph.D. thesis, School of Informatics, University of Edinburgh.

Dagan, Ido, Lillian Lee, and Fernando C. N. Pereira. 1999. Similarity-based models of word co-occurrence probabilities. *Machine Learning*, 34(1):34–69.

Dinu, Georgiana and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-10)*, pages 1,162–1,172, Cambridge, MA.

Erk, Katrin. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 216–223, Prague.

Erk, Katrin and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-08)*, pages 897–906, Honolulu, HI.

Erk, Katrin, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Essen, Ute and Volker Steinbiss. 1992. Co-occurrence smoothing for stochastic language modeling. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*

(ICASSP-92), pages 161–164, San Francisco, CA.

Evert, Stefan. 2004. *The Statistics of Word Co-occurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Fan, Ron-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1,871–1,874.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Finkel, Jenny Rose, Trond Grenager, and Christopher D. Manning. 2007. The infinite tree. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 272–279, Prague.

Frith, J. R. 1957. A Synopsis of Linguistic Theory 1930-1955. In *Studies in Linguistic Analysis*. Oxford Philological Society, Oxford.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian Data Analysis*. Chapman and Hall/ CRC, Boca Raton, FL, 2nd edition.

Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Goldwater, Sharon, Thomas L. Griffiths, and Mark Johnson. 2011. Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, 12:2,335–2,382.

Grefenstette, Edward and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of EMNLP-11*, pages 1,394–1,404, Edinburgh, UK.

Griffiths, Thomas L. and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5,228–5,235.

Griffiths, Thomas L., Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2004. Integrating topics and syntax. In *Proceedings of the 18th Annual Conference on Neural Information Processing Systems (NIPS-04)*, pages 537–544, Vancouver.

Grishman, Ralph and John Sterling. 1993. Smoothing of automatically generated selectional constraints. In *Proceedings of the ARPA Human Language Technology Workshop (HLT-93)*, pages 254–259, Plainsboro, NJ.

Harper, Kenneth E. 1965. Measurement of similarity between nouns. In *Proceedings of the 1965 International Conference on Computational Linguistics (COLING-65)*, New York, NY.

Harris, Zellig. 1954. Distributional structure. *Word*, 10(23):146–162.

Heinrich, Gregor. 2009. Parameter estimation for text analysis. Technical report, Fraunhofer IGD.

Hinton, Geoffrey E. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1,771–1,800.

Katz, Jerrold J. and Jerry A. Fodor. 1963. The stucture of a semantic theory. *Language*, 39(2):170–210.

Keller, Frank and Mirella Lapata. 2003. Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Kishida, Kazuaki. 2005. Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. Technical Report NII-2005-014E, National Institute of Informatics, Tokyo, Japan.

Lee, Lillian. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 25–32, College Park, MD.

Li, Hang and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.

Liang, Percy, Slav Petrov, Michael Jordan, and Dan Klein. 2007. The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP-07)*, pages 688–697, Prague.

McCarthy, Diana and John Carroll. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

McCarthy, Diana and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

McCarthy, Diana, Sriram Venkatapathy, and Aravind K. Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *Proceedings of the 2007 Joint*

Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-07), pages 369–379, Prague.

Meng, Xiao-Li, Robert Rosenthal, and Donald B. Rubin. 1992. Comparing correlated correlation coefficients. *Psychological Bulletin*, 111(1):172–175.

Minka, Thomas P. 2003. Estimating a Dirichlet distribution. Available at `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/`.

Mitchell, Jeff and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, pages 236–244, Columbus, OH.

Mitchell, Jeff and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1,388–1,429.

Moon, Taesun, Katrin Erk, and Jason Baldridge. 2010. Crouching Dirichlet, hidden Markov model: Unsupervised POS tagging with context local tag generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-10)*, pages 196–206, Cambridge, MA.

Ó Séaghdha, Diarmuid. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, pages 435–444, Uppsala.

Ó Séaghdha, Diarmuid and Ann Copestake. 2008. Semantic classification with distributional kernels. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, pages 649–655, Manchester.

Ó Séaghdha, Diarmuid and Anna Korhonen. 2011. Probabilistic models of similarity in syntactic context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP-11)*, pages 1,047–1,057, Edinburgh.

Ó Séaghdha, Diarmuid and Anna Korhonen. 2012. Modelling selectional preferences in a lexical hierarchy. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (*SEM-12)*, pages 170–179, Montreal.

Padó, Sebastian and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Pantel, Patrick, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning inferential selectional preferences. In *Proceedings of NAACL-07*, pages 564–571, Rochester, NY.

Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93)*, pages 183–190, Columbus, OH.

Petrov, Slav, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-06)*, pages 433–440, Sydney.

Rayner, Keith, Tessa Warren, Barbara J. Juhasz, and Simon P. Liversedge. 2004. The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning Memory and Cognition*, 30(6):1,290–1,301.

Reisinger, Joseph and Raymond Mooney. 2010. A mixture model with sharing for lexical semantics. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-10)*, pages 1,173–1,182, Cambridge, MA.

Reisinger, Joseph and Raymond Mooney. 2011. Cross-cutting models of lexical semantics. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP-11)*, pages 1,405–1,415, Edinburgh.

Resnik, Philip. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Ritter, Alan, Mausam, and Oren Etzioni. 2010. A latent Dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, pages 424–434, Uppsala.

Rooth, Mats, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 104–111, College Park, MD.

Russell, Bertrand. 1940. *An Inquiry into Meaning and Truth*. George Allen and Unwin, London.

Schulte im Walde, Sabine, Christian Hying, Christian Scheible, and Helmut Schmid.

2008. Combining EM training and the MDL principle for an automatic verb classification incorporating selectional preferences. In *Proceedings of ACL-08: HLT*, pages 496–504, Columbus, OH.

Shutova, Ekaterina. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-10)*, pages 1,029–1,037, Los Angeles, CA.

Socher, Richard, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS-11)*, pages 801–809, Granada.

Spärck Jones, Karen. 1964. *Synonymy and Semantic Classification*. Ph.D. thesis, University of Cambridge.

Teh, Yee Whye. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-06)*, pages 985–992, Sydney.

Thater, Stefan, Hagen Fürstenau, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, pages 948–957, Uppsala.

Thater, Stefan, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-11)*, pages 1,134–1,143, Hyderabad.

Titov, Ivan and Alexandre Klementiev. 2011. A Bayesian model for unsupervised semantic parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1,445–1,455, Portland, OR.

Turney, Peter D. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44:533–585.

Turney, Peter D. and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Vlachos, Andreas, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and constrained Dirichlet process mixture models for verb clustering. In *Proceedings of the EACL-09 Workshop on Geometrical Models of Natural Language Semantics (GEMS-09)*, pages 74–82, Athens.

Wallach, Hanna, David Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems(NIPS-09)*, pages 1,973–1,981, Vancouver.

Wallach, Hanna M. 2006. Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning (ICML-06)*, pages 977–984, Pittsburgh, PA.

Wallach, Hanna M. 2008. *Structured topic models for language*. Ph.D. thesis, University of Cambridge.

Weeds, Julie and David Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–476.

Wilks, Yorick. 1978. Making preferences more active. *Artificial Intelligence*, 11:197–225.

Yao, Limin, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *Proceedings of EMNLP-11*, pages 1,456–1,466, Edinburgh.

Yao, Limin, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-09)*, pages 937–946, Paris.

Yeh, Alexander. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics (COLING-00)*, pages 947–953, Saarbrücken.

Zapirain, Beñat, Eneko Agirre, and Lluís Màrquez. 2009. Generalizing over lexical features: Selectional preferences for semantic role classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language*

Processing of the AFNLP (ACL-IJCNLP-09), pages 73–76, Singapore.

Zapirain, Beñat, Eneko Agirre, Lluís Màrquez, and Mihai Surdeanu. 2010. Improving semantic role classification with selectional preferences. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-10)*, pages 373–376, Los Angeles, CA.

Zhou, Guangyou, Jun Zhao, Kang Liu, and Li Cai. 2011. Exploiting Web-derived selectional preference to improve statistical dependency parsing. In *Proceedings of ACL-11*, pages 1,556–1,565, Portland, OR.

Zwicky, Arnold M. and Jerrold M. Sadock. 1975. Ambiguity tests and how to fail them. In John P. Kimball, editor, *Syntax and Semantics 4*. Academic Press, New York, NY.