

# Latent Variable Models of Concept-Attribute Attachment

Joseph Reisinger\*

Department of Computer Sciences  
The University of Texas at Austin  
Austin, Texas 78712  
joeraii@cs.utexas.edu

Marius Paşca

Google Inc.  
1600 Amphitheatre Parkway  
Mountain View, California 94043  
mars@google.com

## Abstract

This paper presents a set of Bayesian methods for automatically extending the WORDNET ontology with new concepts and annotating existing concepts with generic property fields, or *attributes*. We base our approach on Latent Dirichlet Allocation and evaluate along two dimensions: (1) the precision of the ranked lists of attributes, and (2) the quality of the attribute assignments to WORDNET concepts. In all cases we find that the principled LDA-based approaches outperform previously proposed heuristic methods, greatly improving the specificity of attributes at each concept.

## 1 Introduction

We present a Bayesian approach for simultaneously extending Is-A hierarchies such as those found in WORDNET (WN) (Fellbaum, 1998) with additional concepts, and annotating the resulting concept graph with attributes, i.e., generic property fields shared by instances of that concept. Examples of attributes include “height” and “eye-color” for the concept *Person* or “gdp” and “president” for *Country*. Identifying and extracting such attributes relative to a set of flat (i.e., non-hierarchically organized) labeled classes of instances has been extensively studied, using a variety of data, e.g., Web search query logs (Paşca and Van Durme, 2008), Web documents (Yoshinaga and Torisawa, 2007), and Wikipedia (Suchanek et al., 2007; Wu and Weld, 2008).

Building on the current state of the art in attribute extraction, we propose a model-based approach for mapping flat sets of attributes annotated with class labels into an existing ontology. This inference problem is divided into two main components: (1) identifying the appropriate parent concept for each labeled class and (2) learning

the correct level of abstraction for each attribute in the extended ontology. For example, consider the task of annotating WN with the labeled class *renaissance painters* containing the class instances Pisanello, Hieronymus Bosch, and Jan van Eyck and associated with the attributes “famous works” and “style.” Since there is no WN concept for *renaissance painters*, the latter would need to be mapped into WN under, e.g., *Painter*. Furthermore, since “famous works” and “style” are not specific to *renaissance painters* (or even the WN concept *Painter*), they should be placed at the most appropriate level of abstraction, e.g., *Artist*. In this paper, we show that both of these goals can be realized jointly using a probabilistic topic model, namely hierarchical Latent Dirichlet Allocation (LDA) (Blei et al., 2003b).

There are three main advantages to using a topic model as the annotation procedure: (1) Unlike hierarchical clustering (Duda et al., 2000), the attribute distribution at a concept node is not composed of the distributions of its children; attributes found specific to the concept *Painter* would not need to appear in the distribution of attributes for *Person*, making the internal distributions at each concept more meaningful as attributes specific to that concept; (2) Since LDA is fully Bayesian, its model semantics allow additional prior information to be included, unlike standard models such as Latent Semantic Analysis (Hofmann, 1999), improving annotation precision; (3) Attributes with multiple related meanings (i.e., polysemous attributes) are modeled implicitly: if an attribute (e.g., “style”) occurs in two separate input classes (e.g., *poets* and *car models*), then that attribute might attach at two different concepts in the ontology, which is better than attaching it at their most specific common ancestor (*Whole*) if that ancestor is too general to be useful. However, there is also a pressure for these two occurrences to attach to a single concept.

We use WORDNET 3.0 as the specific test ontology for our annotation procedure, and evalu-

\*Contributions made during an internship at Google.

**anticancer drugs:** mechanism of action, uses, extravasation, solubility, contraindications, side effects, chemistry, molecular weight, history, mode of action  
**bollywood actors:** biography, filmography, age, bio-data, height, profile, autobiography, new wallpapers, latest photos, family pictures  
**citrus fruits:** nutrition, health benefits, nutritional value, nutritional information, calories, nutrition facts, history  
**european countries:** population, flag, climate, president, economy, geography, currency, population density, topography, vegetation, religion, natural resources  
**london boroughs:** population, taxis, local newspapers, mp, lb, street map, renauld connexions, local history  
**microorganisms:** cell structure, taxonomy, life cycle, reproduction, colony morphology, scientific name, virulence factors, gram stain, clipart  
**renaissance painters:** early life, bibliography, short biography, the david, bio, painting, techniques, homosexuality, birthplace, anatomical drawings, famous paintings

Figure 1: Examples of labeled attribute sets extracted using the method from (Paşca and Van Durme, 2008).

ate three variants: (1) a *fixed structure* approach where each flat class is attached to WN using a simple string-matching heuristic, and concept nodes are annotated using LDA, (2) an extension of LDA allowing for *sense selection* in addition to annotation, and (3) an approach employing a non-parametric prior over tree structures capable of inferring arbitrary ontologies.

The remainder of this paper is organized as follows: §2 describes the full ontology annotation framework, §3 introduces the LDA-based topic models, §4 gives the experimental setup, §5 gives results, §6 gives related work and §7 concludes.

## 2 Ontology Annotation

Input to our ontology annotation procedure consists of sets of class instances (e.g., Pisanello, Hieronymus Bosch) associated with class labels (e.g., *renaissance painters*) and attributes (e.g., “birthplace”, “famous works”, “style” and “early life”). Clusters of noun phrases (instances) are constructed using distributional similarity (Lin and Pantel, 2002; Hearst, 1992) and are labeled by applying “such-as” surface patterns to raw Web text (e.g., “*renaissance painters* such as Hieronymus Bosch”), yielding 870K instances in more than 4500 classes (Paşca and Van Durme, 2008).

Attributes for each flat labeled class are extracted from anonymized Web search query logs using the minimally supervised procedure in (Paşca, 2008)<sup>1</sup>. Candidate attributes are ranked based on their weighted Jaccard similarity to a set of 5 manually provided seed attributes for the

<sup>1</sup>Similar query data, including query strings and frequency counts, is available from, e.g., (Gao et al., 2007)

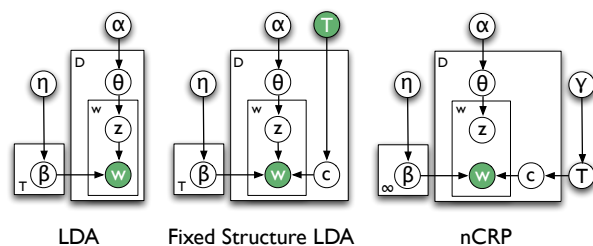


Figure 2: Graphical models for the LDA variants; shaded nodes indicate observed quantities.

class *european countries*. Figure 1 illustrates several such *labeled attribute sets* (the underlying instances are not depicted). Naturally, the attributes extracted are not perfect, e.g., “lb” and “renault connexions” as attributes for *london boroughs*.

We propose a set of Bayesian generative models based on LDA that take as input *labeled attribute sets* generated using an extraction procedure such as the above and organize the attributes in WN according to their level of generality. Annotating WN with attributes proceeds in three steps: (1) attaching labeled attribute sets to leaf concepts in WN using string distance, (2) inferring an attribute model using one of the LDA variants discussed in §3, and (3) generating ranked lists of attributes for each concept using the model probabilities (§4.3).

## 3 Hierarchical Topic Models

### 3.1 Latent Dirichlet Allocation

The underlying mechanism for our annotation procedure is LDA (Blei et al., 2003b), a fully Bayesian extension of probabilistic Latent Semantic Analysis (Hofmann, 1999). Given  $D$  labeled attribute sets  $\mathbf{w}_d$ ,  $d \in D$ , LDA infers an unstructured set of  $T$  latent *annotated concepts* over which attribute sets decompose as mixtures.<sup>2</sup> The latent annotated concepts represent semantically coherent groups of attributes expressed in the data, as shown in Example 1.

The generative model for LDA is given by

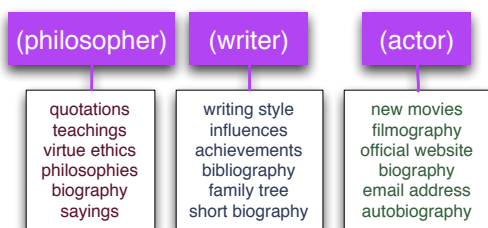
$$\begin{aligned}
 \theta_d | \alpha &\sim \text{Dir}(\alpha), & d \in 1 \dots D \\
 \beta_t | \eta &\sim \text{Dir}(\eta), & t \in 1 \dots T \\
 z_{i,d} | \theta_d &\sim \text{Mult}(\theta_d), & i \in 1 \dots |\mathbf{w}_d| \\
 w_{i,d} | \beta_{z_{i,d}} &\sim \text{Mult}(\beta_{z_{i,d}}), & i \in 1 \dots |\mathbf{w}_d|
 \end{aligned} \tag{1}$$

where  $\alpha$  and  $\eta$  are hyperparameters smoothing the per-attribute set distribution over concepts and per-concept attribute distribution respectively (see Figure 2 for the graphical model). We are interested in the case where  $\mathbf{w}$  is known and we want

<sup>2</sup>In topic modeling literature, attributes are *words* and attribute sets are *documents*.

to compute the conditional posterior of the remaining random variables  $p(\mathbf{z}, \beta, \theta | \mathbf{w})$ . This distribution can be approximated efficiently using Gibbs sampling. See (Blei et al., 2003b) and (Griffiths and Steyvers, 2002) for more details.

**(Example 1)** Given 26 labeled attribute sets falling into three broad semantic categories: philosophers, writers and actors (e.g., sets for *contemporary philosophers*, *women writers*, *bollywood actors*), LDA is able to infer a meaningful set of latent annotated concepts:



(concept labels manually added for the latent annotated concepts are shown in parentheses). Note that with a flat concept structure, attributes can only be separated into broad clusters, so the generality/specificity of attributes cannot be inferred. Parameters were  $\alpha=1, \eta=0.1, T=3$ .

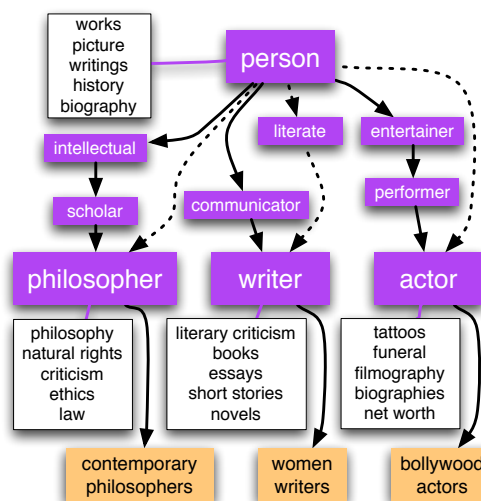
### 3.2 Fixed-Structure LDA

In this paper, we extend LDA to model structural dependencies between latent annotated concepts (cf. (Li and McCallum, 2006; Sivic et al., 2008)); In particular, we fix the concept structure to correspond to the WN Is-A hierarchy. Each labeled attribute set is assigned to a leaf concept in WN based on the edit distance between the concept label and the attribute set label. Possible latent concepts for this set include the concepts along all paths from its attachment point to the WN root, following Is-A relation edges. Therefore, any two labeled attribute sets share a number of latent concepts based on their similarity in WN: all labeled attribute sets share at least the root concept, and may share more concepts depending on their most specific, common ancestor. Under such a model, more general attributes naturally attach to latent concept nodes closer to the root, and more specific attributes attach lower (Example 2).

Formally, we introduce into LDA an extra set of random variables  $\mathbf{c}_d$  identifying the subset of concepts in  $T$  available to attribute set  $d$ , as shown in the diagram at the middle of Figure 2.<sup>3</sup> For example, with a tree structure,  $\mathbf{c}_d$  would be constrained to correspond to the concept nodes in  $T$  on the path from the root to the leaf containing  $d$ . Equation 1 can be adapted to this case if the index  $t$  is taken to range over concepts applicable to attribute set  $d$ .

<sup>3</sup>Abusing notation, we use  $T$  to refer to a structured set of concepts and to refer to the number of concepts in flat LDA

**(Example 2)** Fixing the latent concept structure to correspond to WN (dark/purple nodes), and attaching each labeled attribute set (examples depicted by light/orange nodes) yields the annotated hierarchy:



Attribute distributions for the small nodes are not shown. Dotted lines indicate multiple paths from the root, which can be inferred using sense selection. Unlike with the flat annotated concept structure, with a hierarchical concept structure, attributes can be separated by their generality. Parameters were set at  $\alpha=1$  and  $\eta=0.1$ .

### 3.3 Sense-Selective LDA

For each labeled attribute set, determining the appropriate parent concept in WN is difficult since a single class label may be found in many different synsets (for example, the class *bollywood actors* might attach to the “thespian” sense of *Actor* or the “doer” sense). Fixed-hierarchy LDA can be extended to perform automatic sense selection by placing a distribution over the leaf concepts  $\mathbf{c}$ , describing the prior probability of each possible path through the concept tree. For WN, this amounts to fixing the set of concepts to which a labeled attribute set can attach (e.g., restricting it to a semantically similar subset) and assigning a probability to each concept (e.g., using the relative WN concept frequencies). The probability for each sense attachment  $\mathbf{c}_d$  becomes

$$p(\mathbf{c}_d | \mathbf{w}, \mathbf{c}_{-d}, \mathbf{z}) \propto p(\mathbf{w}_d | \mathbf{c}, \mathbf{w}_{-d}, \mathbf{z}) p(\mathbf{c}_d | \mathbf{c}_{-d}),$$

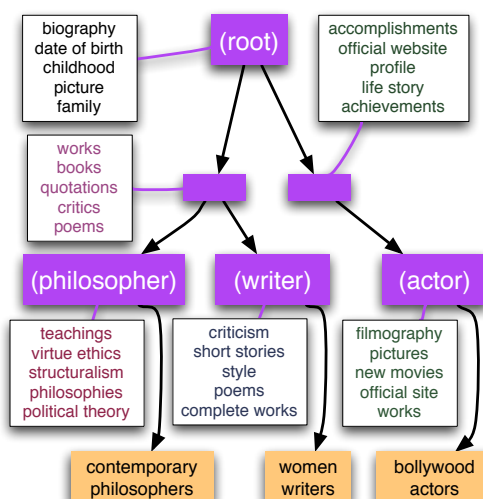
i.e., the complete conditionals for sense selection.  $p(\mathbf{c}_d | \mathbf{c}_{-d})$  is the conditional probability for attaching attribute set  $d$  at  $\mathbf{c}_d$  (e.g., simply the prior  $p(\mathbf{c}_d | \mathbf{c}_{-d}) \stackrel{\text{def}}{=} p(\mathbf{c}_d)$  in the WN case). A closed form expression for  $p(\mathbf{w}_d | \mathbf{c}, \mathbf{w}_{-d}, \mathbf{z})$  is derived in (Blei et al., 2003a).

### 3.4 Nested Chinese Restaurant Process

In the final model, shown in the diagram on the right side of Figure 2, LDA is extended hierarchically to infer *arbitrary* fixed-depth tree structures

from data. Unlike the fixed-structure and sense-selective approaches which use the WN hierarchy directly, the nCRP generates its own annotated hierarchy whose concept nodes do not necessarily correspond to WN concepts (Example 3). Each node in the tree instead corresponds to a latent annotated concept with an arbitrary number of sub-concepts, distributed according to a Dirichlet Process (Ferguson, 1973). Due to its recursive structure, the underlying model is called the *nested Chinese Restaurant Process* (nCRP). The model in Equation 1 is extended with  $c_d|\gamma \sim \text{nCRP}(\gamma, L)$ ,  $d \in D$  i.e., latent concepts for each attribute set are drawn from an nCRP. The hyperparameter  $\gamma$  controls the probability of branching via the per-node Dirichlet Process, and  $L$  is the fixed tree depth. An efficient Gibbs sampling procedure is given in (Blei et al., 2003a).

**(Example 3)** Applying nCRP to the same three semantic categories: philosophers, writers and actors, yields the model:



(manually added labels are shown in parentheses). Unlike in WN, the inferred structure naturally places philosopher and writer under the same subconcept, which is also separate from actor. Hyperparameters were  $\alpha=0.1$ ,  $\eta=0.1$ ,  $\gamma=1.0$ .

## 4 Experimental Setup

### 4.1 Data Analysis

We employ two data sets derived using the procedure in (Paşca and Van Durme, 2008): the *full* set of automatic extractions generated in § 2, and a *subset* consisting of all attribute sets that fall under the hierarchies rooted at the WN concepts *living thing#1* (i.e., the first sense of *living thing*), *substance#7*, *location#1*, *person#1*, *organization#1* and *food#1*, manually selected to cover a high-precision subset of labeled attribute sets. By comparing the results across the two datasets we can

measure each model’s robustness to noise.

In the *full* dataset, there are 4502 input attribute sets with a total of 225K attributes (24K unique), of which 8121 occur only once. The 10 attributes occurring in the most sets (history, definition, picture(s), images, photos, clipart, timeline, clip art, types) account for 6% of the total. For the *subset*, there are 1510 attribute sets with 76K attributes (11K unique), of which 4479 occur only once.

### 4.2 Model Settings

**Baseline:** Each labeled attribute set is mapped to the most common WN concept with the closest label string distance (Paşca, 2008). Attributes are propagated up the tree, attaching to node  $c$  if they are contained in a majority of  $c$ ’s children.

**LDA:** LDA is used to infer a flat set of  $T = 300$  latent annotated concepts describing the data. The concept selection smoothing parameter is set as  $\alpha=100$ . The smoother for the per-concept multinomial over words is set as  $\eta=0.1$ .<sup>4</sup> The effects of concept structure on attribute precision can be isolated by comparing the structured models to LDA.

**Fixed-Structure LDA (fsLDA):** The latent concept hierarchy is fixed based on WN (§ 3.2), and labeled attribute sets are mapped into it as in *baseline*. The concept graph for each labeled attribute set  $w_d$  is decomposed into (possibly overlapping) chains, one for each unique path from the WN root to  $w_d$ ’s attachment point. Each path is assigned a copy  $w_d$ , reducing the bias in attribute sets with many unique ancestor concepts.<sup>5</sup> The final models contain 6566 annotated concepts on average.

**Sense-Selective LDA (ssLDA):** For the sense selective approach (§ 3.3), the set of possible sense attachments for each attribute set is taken to be all WN concepts with the lowest edit distance to its label, and the conditional probability of each sense attachment  $p(c_d)$  is set proportional to its relative frequency. This procedure results in 2 to 3 senses per attribute set on average, yielding models with 7108 annotated concepts.

**Arbitrary hierarchy (nCRP):** For the arbitrary hierarchy model (§ 3.4), we set the maximum tree depth  $L=5$ , per-concept attribute smoother  $\eta=0.05$ , concept assignment smoother  $\alpha=10$  and nCRP branching proportion  $\gamma=1.0$ . The resulting

<sup>4</sup>(Parameter setting) Across all models, the main results in this paper are robust to changes in  $\alpha$ . For nCRP, changes in  $\eta$  and  $\gamma$  affect the size of the learned model but have less effect on the final precision. Larger values for  $L$  give the model more flexibility, but take longer to train.

<sup>5</sup>Reducing the directed-acyclic graph to a tree ontology did not significantly affect precision.

models span 380 annotated concepts on average.

### 4.3 Constructing Ranked Lists of Attributes

Given an inferred model, there are several ways to construct ranked lists of attributes:

**Per-Node Distribution:** In fsLDA and ssLDA, attribute rankings can be constructed directly for each WN concept  $c$ , by computing the likelihood of attribute  $w$  attaching to  $c$ ,  $\mathcal{L}(c|w) = p(w|c)$  averaged over all Gibbs samples (discarding a fixed number of samples for burn-in). Since  $c$ 's attribute distribution is not dependent on the distributions of its children, the resulting distribution is biased towards more specific attributes.

**Class-Entropy (CE):** In all models, the inferred latent annotated concepts can be used to *smooth* the attribute rankings for each labeled attribute set. Each sample from the posterior is composed of two components: (1) a multinomial distribution over a set of WN nodes,  $p(c|\mathbf{w}_d, \alpha)$  for each attribute set  $\mathbf{w}_d$ , where the (discrete) values of  $c$  are WN concepts, and (2) a multinomial distribution over attributes  $p(w|c, \eta)$  for each WN concept  $c$ . To compute an attribute ranking for  $\mathbf{w}_d$ , we have

$$p(w|\mathbf{w}_d) = \sum_c p(w|c, \eta)p(c|\mathbf{w}_d, \alpha).$$

Given this new ranking for each attribute set, we can compute new rankings for each WN concept  $c$  by averaging again over all the  $\mathbf{w}_d$  that appear as (possible indirect) descendants of  $c$ . Thus, this method uses LDA to first perform reranking on the raw extractions before applying the baseline ontology induction procedure (§ 4.2).<sup>6</sup>

CE ranking exhibits a “conservation of entropy” effect, whereby the proportion of general to specific attributes in each attribute set  $\mathbf{w}_d$  remains the same in the posterior. If set  $A$  contains 10 specific attributes and 30 generic ones, then the latter will be favored over the former in the resulting distribution 3 to 1. Conservation of entropy is a strong assumption, and in particular it hinders improving the *specificity* of attribute rankings.

**Class-Entropy+Prior:** The LDA-based models do not inherently make use of any ranking information contained in the original extractions. However, such information can be incorporated in the form of a prior. The final ranking method combines CE with an exponential prior over the attribute rank in the baseline extraction. For each attribute set, we compute the probability of each

<sup>6</sup>One simple extension is to run LDA again on the CE ranked output, yielding an iterative procedure; however, this was not found to significantly affect precision.

attribute  $p(w|\mathbf{w}_d) = p_{\text{lda}}(w|\mathbf{w}_d)p_{\text{base}}(w|\mathbf{w}_d)$ , assuming a parametric form for  $p_{\text{base}}(w|\mathbf{w}_d) \stackrel{\text{def}}{=} \theta^{r(w, \mathbf{w}_d)}$ . Here,  $r(w, \mathbf{w}_d)$  is the rank of  $w$  in attribute set  $d$ . In all experiments reported,  $\theta=0.9$ .

### 4.4 Evaluating Attribute Attachment

For the WN-based models, in addition to measuring the average precision of the reranked attributes, it is also useful to evaluate the assignment of attributes to WN concepts. For this evaluation, human annotators were asked to determine the most appropriate WN synset(s) for a set of gold attributes, taking into account polysemous usage. For each model, ranked lists of possible concept assignments  $C(w)$  are generated for each attribute  $w$ , using  $\mathcal{L}(c|w)$  for ranking. The accuracy of a list  $C(w)$  for an attribute  $w$  is measured by a scoring metric that corresponds to a modification (Paşca and Alfonseca, 2009) of the mean reciprocal rank score (Voorhees and Tice, 2000):

$$DRR = \max_c \frac{1}{\text{rank}(c) \times (1 + \text{PathToGold})}$$

where  $\text{rank}(c)$  is the rank (from 1 up to 10) of a concept  $c$  in  $C(w)$ , and PathToGold is the length of the minimum path along Is-A edges in the conceptual hierarchies between the concept  $c$ , on one hand, and any of the gold-standard concepts manually identified for the attribute  $w$ , on the other hand. The length PathToGold is 0, if the returned concept is the same as the gold-standard concept. Conversely, a gold-standard attribute receives no credit (that is, DRR is 0) if no path is found in the hierarchies between the top 10 concepts of  $C(w)$  and any of the gold-standard concepts, or if  $C(w)$  is empty. The overall precision of a given model is the average of the DRR scores of individual attributes, computed over the gold assignment set (Paşca and Alfonseca, 2009).

## 5 Results

### 5.1 Attribute Precision

Precision was manually evaluated relative to 23 concepts chosen for broad coverage.<sup>7</sup> Table 1 shows precision at  $n$  and the Mean Average Precision (MAP); In all LDA-based models, the Bayes average posterior is taken over all Gibbs samples

<sup>7</sup>(Precision evaluation) Attributes were hand annotated using the procedure in (Paşca and Van Durme, 2008) and numerical precision scores (1.0 for vital, 0.5 for okay and 0.0 for incorrect) were assigned for the top 50 attributes per concept. 25 reference concepts were originally chosen, but 2 were not populated with attributes in any method, and hence were excluded from the comparison.

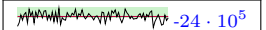





Model	Precision @				MAP
	5	10	20	50	
<b>Base (unranked)</b>	0.45	0.48	0.47	0.44	0.46
<b>Base (ranked)</b>	0.77	0.77	0.69	0.58	0.67
<b>LDA<sup>†</sup></b>					
CE	0.64	0.53	0.52	0.56	0.55
CE+Prior	0.80	0.73	0.74	0.58	0.69
<b>Fixed-structure (fsLDA)</b>					
Per-Node	0.43	0.41	0.42	0.41	0.42
CE	0.75	0.68	0.63	0.55	0.63
CE+Prior	0.78	0.77	0.71	0.59	0.69
<b>Sense-selective (ssLDA)</b>					
Per-Node	0.37	0.44	0.42	0.41	0.42
CE	0.69	0.68	0.65	0.58	0.64
CE+Prior	0.81	0.80	0.72	0.60	0.70
<b>nCRP<sup>†</sup></b>					
CE	0.74	0.76	0.73	0.65	0.72
CE+Prior	0.88	0.85	0.81	0.68	0.78
Subset only					
<b>Base (unranked)</b>	0.61	0.62	0.62	0.60	0.62
<b>Base (ranked)</b>	0.79	0.82	0.72	0.65	0.72
–WN living thing	0.73	0.80	0.71	0.65	0.69
–WN substance	0.80	0.80	0.69	0.53	0.68
–WN location	0.95	0.93	0.84	0.75	0.84
–WN person	0.75	0.83	0.75	0.77	0.77
–WN organization	0.60	0.70	0.60	0.68	0.63
–WN food	0.90	0.85	0.58	0.45	0.64
<b>Fixed-structure (fsLDA)</b>					
Per-Node	0.64	0.58	0.52	0.56	0.55
CE	0.90	0.83	0.78	0.73	0.78
CE+Prior	0.88	0.86	0.80	0.66	0.78
–WN living thing	0.83	0.88	0.78	0.63	0.77
–WN substance	0.85	0.83	0.78	0.66	0.76
–WN location	0.95	0.95	0.88	0.75	0.85
–WN person	1.00	0.93	0.91	0.76	0.87
–WN organization	0.80	0.70	0.80	0.76	0.75
–WN food	0.80	0.70	0.63	0.40	0.59
<b>nCRP<sup>†</sup></b>					
CE	0.88	0.88	0.78	0.71	0.79
CE+Prior	0.90	0.88	0.83	0.67	0.79

Table 1: Precision at  $n$  and mean-average precision for all models and data sets. Inset plots show log-likelihood of each Gibbs sample, indicating convergence except in the case of nCRP. <sup>†</sup> indicates models that do not generate annotated concepts corresponding to WN nodes and hence have no per-node scores.

after burn-in.<sup>8</sup> The improvements in average precision are important, given the amount of noise in the raw extracted data.

When prior attribute rank information (Per-Node and CE scores) from the baseline extractions is *not* incorporated, all LDA-based models outperform the unranked baseline (Table 1). In particular, LDA yields a 17% reduction in error (MAP)

<sup>8</sup>(Bayes average vs. maximum a-posteriori) The full Bayesian average posterior consistently yielded higher precision than the maximum a-posteriori model. For the per-node distributions, the fsLDA Bayes average model exhibits a 17% reduction in relative error over the maximum a-posteriori estimate and for ssLDA there was a 26% reduction.

Model	DRR Scores			
	all	(n)	found	(n)
<b>Base (unranked)</b>	0.14	(150)	0.24	(91)
<b>Base (ranked)</b>	0.17	(150)	0.21	(123)
<b>Fixed-structure (fsLDA)</b>	0.31	(150)	0.37	(128)
<b>Sense-selective (ssLDA)</b>	0.31	(150)	0.37	(128)
Subset only				
<b>Base (unranked)</b>	0.15	(97)	0.27	(54)
<b>Base (ranked)</b>	0.18	(97)	0.24	(74)
WN living thing	0.29	(27)	0.35	(22)
WN substance	0.21	(12)	0.32	(8)
WN location	0.12	(30)	0.17	(20)
WN person	0.37	(18)	0.44	(15)
WN organization	0.15	(31)	0.17	(27)
WN food	0.15	(6)	0.22	(4)
<b>Fixed-structure (fsLDA)</b>	0.37	(97)	0.47	(77)
WN living thing	0.45	(27)	0.55	(22)
WN substance	0.48	(12)	0.64	(9)
WN location	0.34	(30)	0.44	(23)
WN person	0.44	(18)	0.52	(15)
WN organization	0.44	(31)	0.71	(19)
WN food	0.60	(6)	0.72	(5)

Table 2: All measures the DRR score relative to the entire gold assignment set; *found* measures DRR only for attributes with  $DRR(w) > 0$ ;  $n$  is the number of scores averaged.

over the baseline, fsLDA yields a 31% reduction, ssLDA yields a 33% reduction and nCRP yields a 48% reduction (24% reduction over fsLDA). Performance also improves relative to the *ranked* baseline when prior ranking information is incorporated in the LDA-based models, as indicated by CE+Prior scores in Table 1. LDA and fsLDA reduce relative error by 6%, ssLDA by 9% and nCRP by 33%. Furthermore, nCRP precision *without* ranking information surpasses the baseline with ranking information, indicating robustness to extraction noise. Precision curves for individual attribute sets are shown in Figure 3. Overall, learning unconstrained hierarchies (nCRP) increases precision, but as the inferred node distributions do not correspond to WN concepts they cannot be used for annotation.

One benefit to using an admixture model like LDA is that each concept node in the resulting model contains a distribution over attributes specific only to that node (in contrast to, e.g., hierarchical agglomerative clustering). Although absolute precision is lower as more general attributes have higher average precision (Per-Node scores in Table 1), these distributions are semantically meaningful in many cases (Figure 4) and furthermore can be used to calculate concept assignment precision for each attribute.<sup>9</sup>

<sup>9</sup>Per-node distributions (and hence DRR) were not evalu-

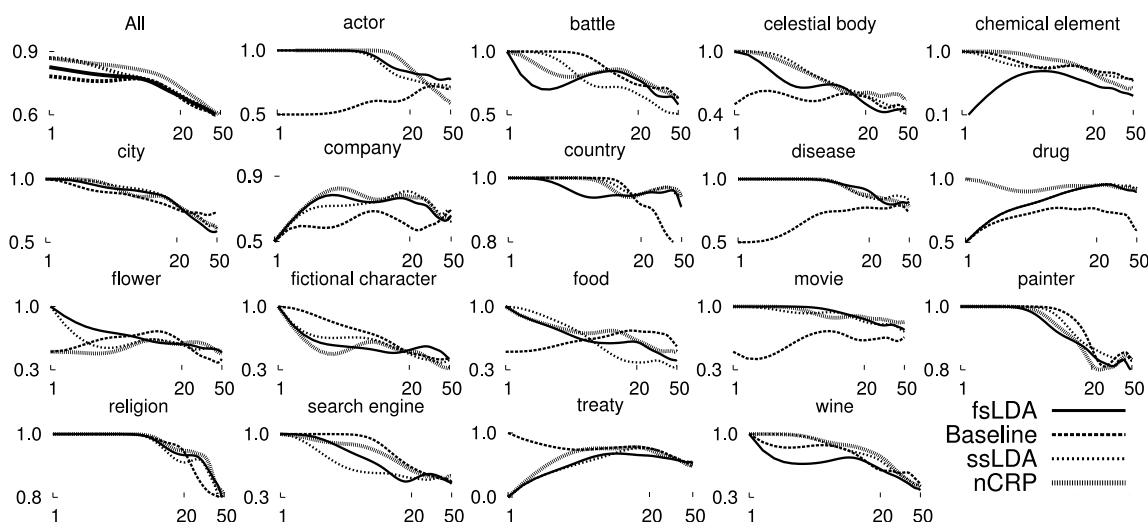


Figure 3: Precision (%) vs. rank plots (log scale) of attributes broken down across 18 labeled test attribute sets. Ranked lists of attributes are generated using the CE+Prior method.

## 5.2 Concept Assignment Precision

The precision of assigning attributes to various concepts is summarized in Table 2. Two scores are given: *all* measures DRR relative to the entire gold assignment set, and *found* measures DRR only for attributes with  $DRR(w) > 0$ . Comparing the scores gives an estimate of whether coverage or precision is responsible for differences in scores. fsLDA and ssLDA both yield a 20% reduction in relative error (17.2% increase in absolute DRR) over the unranked baseline and a 17.2% reduction (14.2% absolute increase) over the ranked baseline.

## 5.3 Subset Precision and DRR

Precision scores for the manually selected subset of extractions are given in the second half of Table 1. Relative to the unranked baseline, fsLDA and nCRP yield 42% and 44% reductions in error respectively, and relative to the ranked baseline they both yield a 21.4% reduction. In terms of absolute precision, there is no benefit to adding in prior ranking knowledge to fsLDA or nCRP, indicating diminishing returns as average baseline precision increases (Baseline vs. fsLDA/nCRP CE scores). Broken down across each of the subhierarchies, LDA helps in all cases except *food*.

DRR scores for the subset are given in the lower half of Table 2. Averaged over all gold test attributes, DRR scores double when using fsLDA. These results can be misleading, however, due to artificially low coverage. Hence, Table 2 also shows DRR scores broken down over each subhierarchy. In this case fsLDA more than doubles the DRR relative to the baseline for *substance* and *location*, and triples it for *organization* and *food*.

ated for LDA or nCRP, because they are not mapped to WN.

## 6 Related Work

A large body of previous work exists on extending WORDNET with additional concepts and instances (Snow et al., 2006; Suchanek et al., 2007); these methods do not address attributes directly. Previous literature in attribute extraction takes advantage of a range of data sources and extraction procedures (Chklovski and Gil, 2005; Tokunaga et al., 2005; Paşca and Van Durme, 2008; Yoshinaga and Torisawa, 2007; Probst et al., 2007; Van Durme et al., 2008; Wu and Weld, 2008). However these methods do not address the task of determining the level of specificity for each attribute. The closest studies to ours are (Paşca, 2008), implemented as the baseline method in this paper; and (Paşca and Alfonseca, 2009), which relies on heuristics rather than formal models to estimate the specificity of each attribute.

## 7 Conclusion

This paper introduced a set of methods based on Latent Dirichlet Allocation (LDA) for jointly extending the WORDNET ontology and annotating its concepts with attributes (see Figure 4 for the end result). LDA significantly outperformed a previous approach both in terms of the concept assignment precision (i.e., determining the correct level of generality for an attribute) and the mean-average precision of attribute lists at each concept (i.e., filtering out noisy attributes from the base extraction set). Also, relative precision of the attachment models was shown to improve significantly when the raw extraction quality increased, showing the long-term viability of the approach.





## References

- D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. 2003a. Hierarchical topic models and the nested Chinese restaurant process. In *Proceedings of the 17th Conference on Neural Information Processing Systems (NIPS-2003)*, pages 17–24, Vancouver, British Columbia.
- D. Blei, A. Ng, and M. Jordan. 2003b. Latent dirichlet allocation. *Machine Learning Research*, 3:993–1022.
- T. Chklovski and Y. Gil. 2005. An analysis of knowledge collected from volunteer contributors. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*, pages 564–571, Pittsburgh, Pennsylvania.
- R. Duda, P. Hart, and D. Stork. 2000. *Pattern Classification*. John Wiley and Sons.
- C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press.
- T. Ferguson. 1973. A bayesian analysis of some non-parametric problems. *Annals of Statistics*, 1(2):209–230.
- W. Gao, C. Niu, J. Nie, M. Zhou, J. Hu, K. Wong, and H. Hon. 2007. Cross-lingual query suggestion using query logs of different languages. In *Proceedings of the 30th ACM Conference on Research and Development in Information Retrieval (SIGIR-07)*, pages 463–470, Amsterdam, The Netherlands.
- T. Griffiths and M. Steyvers. 2002. A probabilistic approach to semantic representation. In *Proceedings of the 24th Conference of the Cognitive Science Society (CogSci02)*, pages 381–386, Fairfax, Virginia.
- M. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 539–545, Nantes, France.
- T. Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR-99)*, pages 50–57, Berkeley, California.
- W. Li and A. McCallum. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning (ICML-06)*, pages 577–584, Pittsburgh, Pennsylvania.
- D. Lin and P. Pantel. 2002. Concept discovery from text. In *Proceedings of the 19th International Conference on Computational linguistics (COLING-02)*, pages 1–7, Taipei, Taiwan.
- M. Paşca and E. Alfonseca. 2009. Web-derived resources for Web Information Retrieval: From conceptual hierarchies to attribute hierarchies. In *Proceedings of the 32nd International Conference on Research and Development in Information Retrieval (SIGIR-09)*, Boston, Massachusetts.
- M. Paşca and B. Van Durme. 2008. Weakly-supervised acquisition of open-domain classes and class attributes from web documents and query logs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, pages 19–27, Columbus, Ohio.
- M. Paşca. 2008. Turning Web text and search queries into factual knowledge: Hierarchical class attribute extraction. In *Proceedings of the 23rd National Conference on Artificial Intelligence (AAAI-08)*, pages 1225–1230, Chicago, Illinois.
- K. Probst, R. Ghani, M. Krema, A. Fano, and Y. Liu. 2007. Semi-supervised learning of attribute-value pairs from product descriptions. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 2838–2843, Hyderabad, India.
- J. Sivic, B. Russell, A. Zisserman, W. Freeman, and A. Efros. 2008. Unsupervised discovery of visual object class hierarchies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-08)*, pages 1–8, Anchorage, Alaska.
- R. Snow, D. Jurafsky, and A. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-06)*, pages 801–808, Sydney, Australia.
- F. Suchanek, G. Kasneci, and G. Weikum. 2007. Yago: a core of semantic knowledge unifying WordNet and Wikipedia. In *Proceedings of the 16th World Wide Web Conference (WWW-07)*, pages 697–706, Banff, Canada.
- K. Tokunaga, J. Kazama, and K. Torisawa. 2005. Automatic discovery of attribute words from Web documents. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages 106–118, Jeju Island, Korea.
- B. Van Durme, T. Qian, and L. Schubert. 2008. Class-driven attribute extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*, pages 921–928, Manchester, United Kingdom.
- E.M. Voorhees and D.M. Tice. 2000. Building a question-answering test collection. In *Proceedings of the 23rd International Conference on Research and Development in Information Retrieval (SIGIR-00)*, pages 200–207, Athens, Greece.
- F. Wu and D. Weld. 2008. Automatically refining the Wikipedia infobox ontology. In *Proceedings of the 17th World Wide Web Conference (WWW-08)*, pages 635–644, Beijing, China.
- N. Yoshinaga and K. Torisawa. 2007. Open-domain attribute-value acquisition from semi-structured texts. In *Proceedings of the 6th International Semantic Web Conference (ISWC-07), Workshop on Text to Knowledge: The Lexicon/Ontology Interface (OntoLex-2007)*, pages 55–66, Busan, South Korea.