Modeling Concept-Attribute Structure

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

We apply hierarchical Latent Dirichlet Allocation (hLDA) to the problem of *ontology annotation*; automatically extending WORDNET with new concepts and annotating existing concepts with generic property fields, or *attributes*. The resulting annotations are evaluated along two dimensions: (1) the precision of the ranked lists of attributes at each concept, and (2) the specificity of the attribute assignments to WORDNET concepts. We find that hLDA and several variants outperform previously proposed heuristic methods, significantly improving the resulting annotation quality, and confirming the ability of topic models to reduce extraction noise.

1 Introduction

We present a Bayesian approach for simultaneously extending Is-A hierarchies such as those found in WORDNET (WN) [3] 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 [7], Web documents [10], and Wikipedia [8, 9].

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*. Both of these goals can be realized jointly using a probabilistic topic model, namely hierarchical Latent Dirichlet Allocation (hLDA) [1].

There are three main advantages to using a topic model as the annotation procedure: (1) Unlike hierarchical clustering [2], 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 related models such as Latent Semantic Analysis [5], 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

[*] 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, biodata, height, profile, autobiography, new wallpapers, latest photos, family pictures
citrus fruits: nutrition, health benefits, nutritional value, nutritional information, calories, nutrition facts
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, renault 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 [7].

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.

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 [6, 4] and are labeled by applying "such-as" surface patterns to raw Web text (e.g., "*renaissance painters* such as Hieronymous Bosch"), yielding 870K instances in more than 4500 classes [7].

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 below, and (3) generating ranked lists of attributes for each concept using the model probabilities.

We apply three LDA-variants to the concept annotation problem:

1. a *fixed structure* approach (fsLDA) where each flat class is attached to WN using a simple string-matching heuristic, and concept nodes along its hypernym graph are annotated using LDA;
2. an extension of LDA allowing for *sense selection* (ssLDA) in addition to annotation. ssLDA places a distribution over possible attachment points in WN for each flat class, based on a prior distribution over label edit-distance;
3. an approach employing the nested Chinese Restaurant Process (nCRP), a nonparametric prior over tree structures capable of inferring arbitrary ontologies [1]. 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. Thus, internal nodes in the hierarchy do not have labels, but the resulting inferred structure can still be used to smooth the extracted attribute distributions, leading to significant improvements in precision.

3 Results

All three methods significantly outperformed vanilla LDA, as well as a previous heuristic approach, both in terms of the concept assignment precision (i.e., determining the correct level of generality for an attribute; Table 1) and the mean-average precision of attribute lists at each concept (i.e., filtering out noisy attributes from the base extraction set; Table 2) based on an extensive set of human evaluations. In general, ssLDA exhibited only a small improvement over fsLDA, but the benefits are more when prior information about sense frequency is taken into account. The nCRP significantly outperforms both methods relying on the WN ontology when used as a smoother, but does not yield

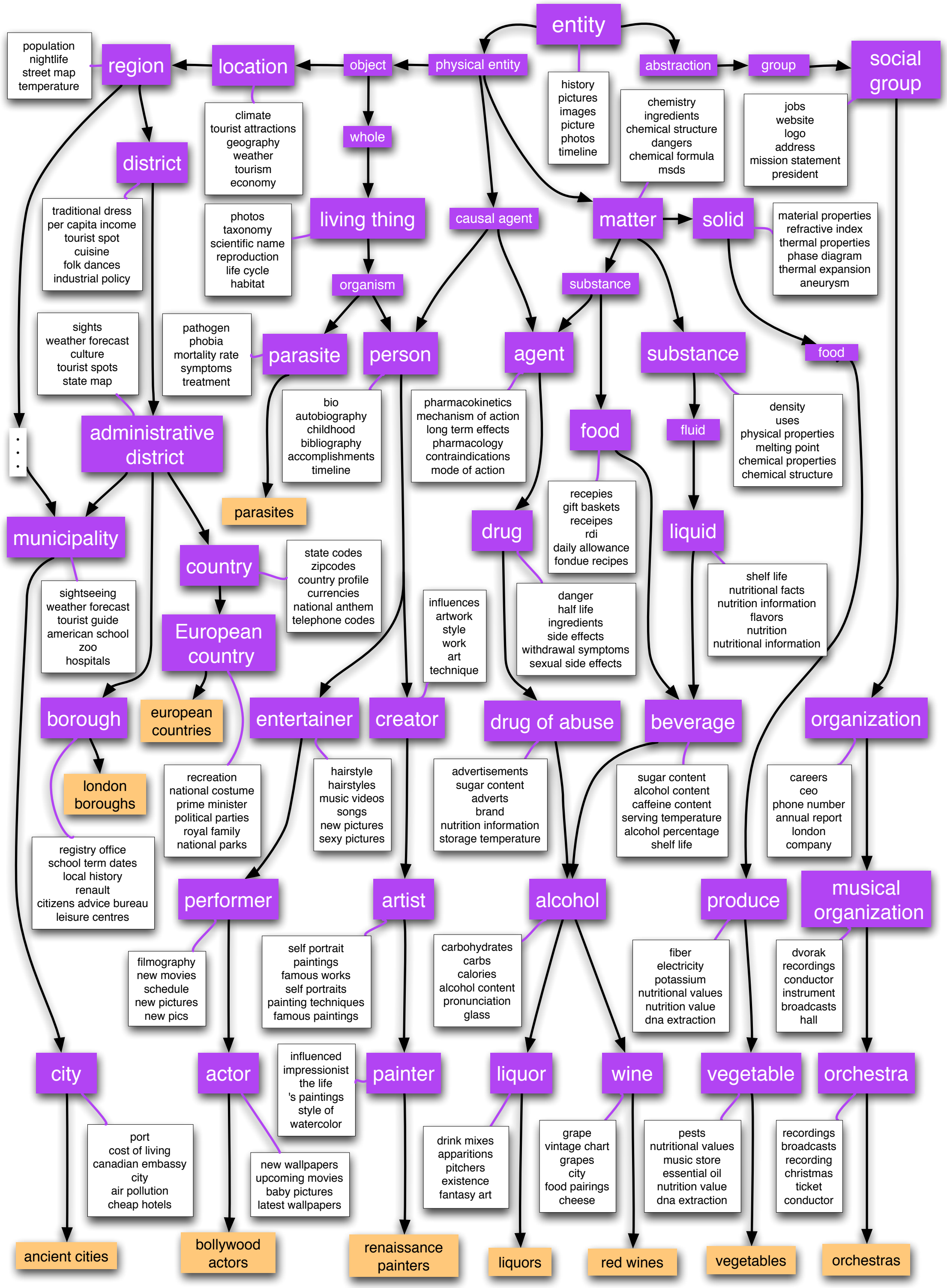

Figure 2: Example per-node attribute distribution generated by fsLDA. Light/orange nodes represent labeled attribute sets attached to WN, and the full hypernym graph is given for each in dark/purple nodes. White nodes depict the top attributes predicted for each WN concept. These inferred annotations exhibit a high degree of concept specificity, naturally becoming more general at higher levels of the ontology. Some annotations, such as for the concepts *Agent*, *Substance*, *Living Thing* and *Person* have high precision and specificity while others, such as *Liquor* and *Actor* need improvement. Overall, the more general concepts yield better annotations as they are averaged over many labeled attribute sets, reducing noise.

Model	Attachment Quality (DRR)			
	all	(n)	found	(n)
Base (unranked)	0.14	(150)	0.24	(91)
Base (ranked)	0.17	(150)	0.21	(123)
Fixed-structure (fsLDA)	0.31	(150)	0.37	(128)
Sense-selective (ssLDA)	0.31	(150)	0.37	(128)

Table 1: Attachment quality; defined as $DRR \overset{\text{def}}{=} \max[rank(c) \times (1 + PathToGold)]^{-1}$where $rank(c)$ is the inferred rank of a concept c for attribute w, and PathToGold is the length of the minimum WordNet distance between the concept c and the gold-standard concepts for the attribute w. *All* measures the attribute-to-concept attachment quality relative to the entire gold assignment set. *found* measures DRR only for attributes with DRR(w)>0; n is the number of scores.

Model	Precision @				MAP
	5	10	20	50	
Baseline	0.45	0.48	0.47	0.44	0.46
+Prior	0.77	0.77	0.69	0.58	0.67
LDA	0.64	0.53	0.52	0.56	0.55
+Prior	0.80	0.73	0.74	0.58	0.69
Fixed-structure (fsLDA)	0.75	0.68	0.63	0.55	0.63
+Prior	0.78	0.77	0.71	0.59	0.69
Sense-selective (ssLDA)	0.69	0.68	0.65	0.58	0.64
+Prior	0.81	0.80	0.72	0.60	0.70
nCRP	0.74	0.76	0.73	0.65	0.72
+Prior	0.88	0.85	0.81	0.68	0.78

Table 2: Precision at n and mean-average precision for all models. *+Prior* indicates models that were trained using additional prior information about attribute ranking.

labeled intermediate concept nodes. Figure 2 shows a small portion of the final annotated hierarchy inferred by fsLDA.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Machine Learning Research*, 3:993–1022, 2003.

[2] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, 2000.

[3] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press, 1998.

[4] M. Hearst. 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, 1992.

[5] T. Hofmann. 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, 1999.

[6] D. Lin and P. Pantel. Concept discovery from text. In *Proceedings of the 19th International Conference on Computational linguistics (COLING-02)*, pages 1–7, Taipei, Taiwan, 2002.

[7] M. Paşca and B. Van Durme. 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, 2008.

[8] F. Suchanek, G. Kasneci, and G. Weikum. 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, 2007.

[9] F. Wu and D. Weld. Automatically refining the Wikipedia infobox ontology. In *Proceedings of the 17th World Wide Web Conference (WWW-08)*, pages 635–644, Beijing, China, 2008.

[10] N. Yoshinaga and K. Torisawa. Open-domain attribute-value acquisition from semi-structured texts. In *Proc. of the 6th International Semantic Web Conference (ISWC-07), Workshop on Text to Knowledge: The Lexicon/Ontology Interface (OntoLex-2007)*, pages 55–66, 2007.