Discovering Subsumption Relationships for Web-Based Ontologies

Dana Movshovitz-Attias[†] Steven Euijong Whang[§] Natalya Noy[§] Alon Halevy[§]

[†]Carnegie Mellon University
dma@cs.cmu.edu

[§]Google Research
{swhang, noy, halevy}@google.com

ABSTRACT

As search engines are becoming smarter at interpreting user queries and providing meaningful responses, they rely on ontologies to understand the meaning of entities. Creating ontologies manually is a laborious process, and resulting ontologies may not reflect the way users think about the world, as many concepts used in queries are noisy, and not easily amenable to formal modeling. There has been considerable effort in generating ontologies from Web text and query streams, which may be more reflective of how users query and write content. In this paper, we describe the LATTE system that automatically generates a subconcept–superconcept hierarchy, which is critical for using ontologies to answer queries. LATTE combines signals based on word-vector representations of concepts and dependency parse trees; however, LATTE derives most of its power from an ontology of attributes extracted from the Web that indicates the aspects of concepts that users find important. LATTE achieves an F1 score of 74%, which is comparable to expert agreement on a similar task. We additionally demonstrate the usefulness of LATTE in detecting high quality concepts from an existing resource of IsA links.

1. INTRODUCTION

One of the recent dramatic changes in Web search is the appearance of answers in response to user queries, complementing the usual collection of Web links. If a user searches for a politician or a movie, all major search engines display a "knowledge panel" with salient information about the entity in the query, including a collection of related entities, or entities of a similar "type" (e.g., movies with the same actor). These answers are derived from large knowledge bases modeled by ontologies—structured formal or semi-formal descriptions of entities and their attributes. While ontologies and knowledge bases have been essential components of intelligent systems for decades, the size and coverage of ontologies used by search engines today is largely unprecedented. Google's Knowledge Graph, for example, is reported to have more than 570 million entities [4].

Even with enormous resources spent on developing knowledge bases, they still cover only a fraction of the concepts, instances, and attributes that users are searching for. As a consequence, many queries that can be answered by structured data are not recognized as such by search engines. In response, several approaches have been used for mining the long tail of concepts and attributes that interest users. Hearst patterns [11] are commonly used to extract IsA relations, for example, the text "Asian countries, such as China," indicates China is an instance of *Asian countries*. The resulting IsA relations can be either subconcept–superconcept or instance–concept relations. Along the same lines, previous work extracted an order of magnitude more attributes than is modeled in Freebase by mining a query stream and Web text [10] (e.g., discovering that countries may have COFFEE PRODUCTION or RAILWAY MINISTER). These signals display broad coverage, however they contain noise.

A crucial aspect of a useful ontology is having an accurate subsumption hierarchy between its concepts. For example, the ontology should know that EUROPEAN CAPITALS is a subconcept of EUROPEAN CITIES. Creating such a hierarchy is challenging for two reasons. First, it turns out that the logical definition of subsumption (i.e., concept A is a subset of B if and only if every instance of A is an instance of B) is limiting in the context of Web ontologies, and many subsumption relations are accepted in practice, though they do not satisfy this definition (see Section 2). The second challenge is that traditional signals do not produce a concept hierarchy with high precision. Specifically, analyzing dependency parse trees of concepts, their word-vector representations [16], or distributional similarity [7, 22] is not sufficient to determine that one is a subconcept of the other.

This paper describes the LATTE[1] system that produces a subsumption hierarchy for concepts extracted from the Web. The observation underlying LATTE is that attributes that are frequently associated with instances of concepts are indicative of subsumption relationships between them. LATTE uses this signal in conjunction with semantic and distributional features to produce a hierarchy that corresponds to commonly accepted subsumptions. Examples of extracted relations are shown in Figure 1. Our experiments demonstrate that LATTE achieves an F1 score of 74%, which is comparable to expert agreement on a similar task. Finally, we show that using predictions made by LATTE we can detect high-quality concepts from a noisy resource of IsA links.

[1]LATTE is a popular espresso-based drink whose interpretation varies significantly depending on location and context.

WebDB'15, May 31 - June 04 2015, Melbourne, VIC, Australia
 ISBN 978-1-4503-3627-7/15/05 $15.00
DOI: http://dx.doi.org/10.1145/2767109.2767111

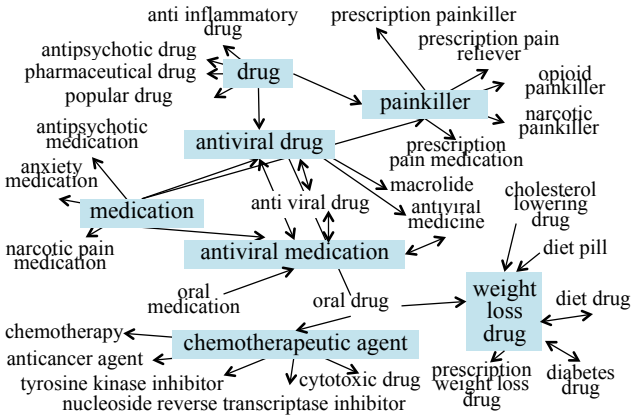

Figure 1: Subsumption relations predicted with high probability. For example, DRUG was predicted as a superconcept of PAINKILLER; it was predicted that WEIGHT LOSS DRUG and DIET DRUG were synonyms, by finding bidirectional subsumption between them.

Category	Subconcept	Superconcept
Clear	PHOTO	VISUAL MATERIAL
	PART TIME JOB	WORK
Ambiguous concept names	PEN	SCHOOL ITEM
	STREAM	NATURAL OBSTACLE
Subjective or relative relations	MONKEY	LARGE ANIMAL
	GLUTEN	HEALTH FOOD
Religious/political views	POPE	POLITICIAN
Temporal and local	CELLULAR PHONE	SMALL DEVICE
	METAL HYBRID	SOLID
Large overlap	LANTHANIDES	F-BLOCK
Common references	LED LIGHT	TECHNOLOGY
Generic superconcepts	ENTREE	CATEGORY

Table 1: Subsumption relations found in our data.

2. SUBSUMPTION IN WEB ONTOLOGIES

Given a set of concepts $\mathcal{C}$, our goal is to discover subsumption relationships among pairs of concepts in $\mathcal{C}$. For example, a UNIVERSITY is a type of organization that has a physical location, so we would like to discover that it is a subconcept of ORGANIZATION and LOCATION. We denote that concept c_1 is a subconcept of c_2 by $c_1 \sqsubseteq c_2$. Subsumption is directional: if $c_1 \sqsubseteq c_2$ then $c_2 \not\sqsubseteq c_1$ unless c_1 and c_2 are synonyms. Note that subsumption is different from the IsA relationship that typically holds between an *instance* and a concept (e.g., STOCKHOLM IsA CITY). Subsumption hierarchies add significant power to ontologies in Web search because many of the attributes of an entity (concept or instance) are attached to its superconcepts, therefore it is important to be able to navigate the hierarchy of concepts [8, 31, 2].

The definition of subsumption in Mathematical Logic requires that $c_1 \sqsubseteq c_2$ if and only if every instance of c_1 is an instance of c_2. However, in the context of Web-based ontologies, this definition is too limiting as it deprives us from considering diverse and context-dependent aspects of the world. Table 1 illustrates the broader types of subsumption we encounter in practice. The first examples show clear subconcept–superconcept relations (PHOTO $\sqsubseteq$ VISUAL MATERIAL). In the following examples, concept names are ambiguous or lack context: a PEN is an item used in school, but also an enclosure in which animals are kept. Some relations are subjective, or depend on a relative scale, or a religious or political affiliation: some MONKEYs are LARGE ANIMALs, others are small, and some are just larger than others. Similarly, some relations hold only temporally or locally: METAL HYBRIDs can be SOLID, depending on the environmental conditions. In some cases, the subsumption is almost complete: all but one of the LANTHANIDE chemical elements belong in the F-BLOCK. In some examples, the common way of referring to a concept is not accurate: LED is a technology that produces light sources with high luminous efficacy. The LED LIGHT itself is a light source, but it is widely used to refer to the LED technology, making it a good subconcept of TECHNOLOGY. Finally, some concepts are so general that they do not contribute to a useful hierarchy (e.g., CATEGORY, TERM). In the manual evaluation of LATTE, we instructed evaluators to examine relations based on whether there is a reasonable scenario in which they hold. We additionally assessed the agreement among evaluators.

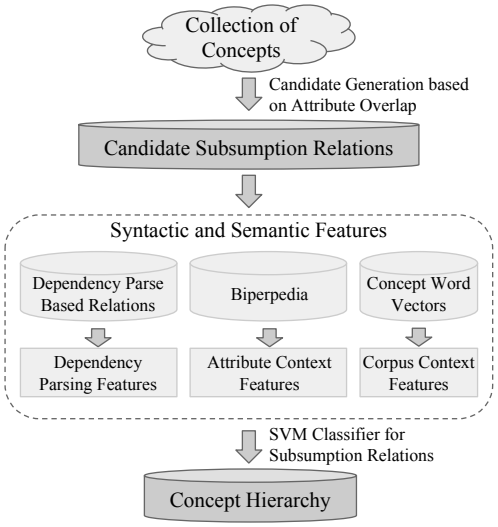

Figure 2: The LATTE system pipeline.

3. THE LATTE SYSTEM

Figure 2 shows the LATTE system pipeline for building a concept hierarchy. The pipeline takes an input set of concepts $\mathcal{C}$. The key stages of the pipeline are: (1) We generate *candidate subsumption relations* from input concepts. (2) We generate features representing contextual and semantic properties of concepts in candidate relations. (3) We predict subsumption relations between concepts using an SVM classifier (we use the LibSVM [5] linear kernel classifier).

The key to the performance of LATTE lies in the features we attach to each concept. Apriori, we imagined that we could leverage state of the art techniques in Natural Language Processing, such as the dependency parse of concept names, and advances in Deep Learning using *word vectors* [16, 18], which represent the local context of a concept in a text corpus. However, we found that these methods do not produce good subsumption relations. A stronger signal

Concept: University		
Academic/Student Life	**Organization**	**Location**
Football roster	Online payment	Country
Admission statistics	Non profit	Address
Sororities	Tax return	Zip code

Table 2: Sample attributes of University.

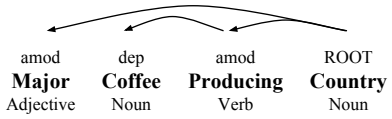

Figure 3: Dependency parse tree for the concept Major coffee producing country.

comes from examining the attributes that are associated (in queries and Web text) with instances of the concepts.

Latte builds on techniques that extract attributes for concepts [20, 1, 10]. In particular, Latte uses Biperpedia [10], which mines thousands of attributes for every concept from search queries and Web text. In Biperpedia, attributes represent binary relationships between two entities (e.g., ceo) or between an entity and a value (e.g., gdp). The concept University, for example, has 2,787 attributes (some listed in Table 2, grouped by topics). Some address academic or student life (e.g., Sororities). Others reflect the fact that a university is a Non profit organization, allowing Online payment. Additional attributes recognize that universities have a physical location, with an Address and Zip code. In Biperpedia, the organizational attributes of University are also attributes of the concept Organization and the location attributes are attributes of Location. It seems plausible, that attributes shared by two concepts give strong evidence of whether a subsumption holds between them. We now describe the components of Latte.

3.1 Candidate Subsumption Relations

If we consider all concepts mentioned on the Web, a randomly selected pair of concepts is not likely to be related. It is also not computationally feasible to examine all pairs in search for subsumption relations. We produce candidate pairs that are likely related, based on their common attributes, and then evaluate a subsumption relation between them. We create two indexes: (1) for an attribute a, C_a is the set of concepts that have this attribute; (2) C'_a contains only the concepts for which attribute a is one of their top ranking attributes, using a TF-IDF based ranking (see Section 3.2.3). The Cartesian product of C_a and C'_a contains pairs of concepts that share at least attribute a and it is important for at least one of them. We use only C_a sets with fewer than 5,000 concepts, and eliminate attributes that are too frequent. For each concept pair $\langle c_1, c_2 \rangle \in C_a \times C'_a$, where $c_1 \neq c_2$, we evaluate in the rest of the pipeline subsumption relations in both directions: $c_1 \sqsubseteq c_2$ and $c_2 \sqsubseteq c_1$.

3.2 Computing Features for Concepts

For each candidate subsumption relation $c_1 \sqsubseteq c_2$, we generate features that reflect similarity, or a directional relationship, between c_1 and c_2. We first explore two baseline approaches to generating features that indicate subsumption: a rule-based method using the dependency parse tree of a concept name, and using word vectors to assess the contextual similarity of concepts. Finally, we evaluate contextual similarity based on the attributes of c_1 and c_2.

3.2.1 Dependency Parsing

As a baseline, we consider an intuitive semantic analysis of concept names. Given the dependency parse tree of a string representing concept $c \in \mathcal{C}$ (e.g., Figure 3), we follow linear paths on the dependency edges starting from the ROOT noun and produce partial terms, which we treat as potential superconcepts to c. For the concept in Figure 3, the extracted superconcepts are then, Country, Major country, Producing country and Coffee producing country. We denote a dependency-parse based subsumption by $\sqsubseteq_{dp}$. We keep relations whose concepts both appear in $\mathcal{C}$, so if Producing country $\notin \mathcal{C}$, we remove dependency based subsumptions that include this concept. Note that this method does not always produce reliable relations, e.g., a Hot dog is not a subconcept of Dog. Dependency parsers are also prone to errors, especially on complex noun phrases.

Given a candidate subsumption relation $c_1 \sqsubseteq c_2$ (created following the procedure described in Section 3.1), we create two binary features that indicate whether the *dependency-based* relation $c_1 \sqsubseteq_{dp} c_2$, or the *inverse* relation $c_2 \sqsubseteq_{dp} c_1$ appear in the set of generated dependency-parse relations (giving two *true/false* features). Intuitively, we expect the former feature to provide positive evidence for the existence of a subsumption relation, and the latter negative evidence.

3.2.2 Word Vectors

Models based on word-vector representations successfully detect continuous similarities among words, normally estimated by comparing the distance or angle of word vector encodings in high-dimensional space. The vectors contain a dense representation of local word context in a large corpus, and they are trained using a Neural Net Language Model [16, 18]. Recently, word vectors have been achieving state-of-the-art performance on a variety of pattern-recognition tasks including machine translation, sentiment analysis, and word sense disambiguation [17, 32, 26, 13, 28]. Many of these tasks rely on the fact that the vector representations allow for a quick and reliable way to assess syntactic and semantic similarities between entities, an observation that is also useful for recovering subsumption relations.

We trained 300-dimensional word vectors on a corpus sampled from Google News (containing 16B words), and we use them to assess similarities between concepts. We treat concept names as compound words, meaning that we produce a single vector representation for each concept. Given a candidate relation $c_1 \sqsubseteq c_2$, with word vectors v_1 and v_2 for the respective concepts, we evaluate their similarity using several distance functions. Cosine distance is commonly used to assess word vector similarity. We additionally use the symmetric Chebychev and Euclidean distances. Finally, we assess a *directional* relation between c_1 and c_2 with an asymmetric variation on Euclidean distance: $\sum_i (v_{1_i} - v_{2_i})$. These metrics constitute our features based on word vectors.

3.2.3 Attributes of Concepts

We now detail features based on attributes in Biperpedia.

Relationships between attribute sets: We consider the intersection and difference in attributes associated with concepts in a candidate relation $c_1 \sqsubseteq c_2$. Consider the overlap among attributes of the concepts University and Lo-

CATION. The attributes COUNTRY and MAP are in the intersection of the attribute groups; intuitively, they account for the similarities among the concepts. However, WIND DIRECTION is only an attribute of LOCATION and SORORITIES is only an attribute of UNIVERSITY so, intuitively, they account for the specialization of either concept relative to the other. We describe the relationship between A_1 and A_2, the attribute sets of c_1 and c_2, by the size of the intersection, union, Jaccard similarity, the relative size of the intersection to A_1 and A_2, and the absolute size of each attribute set.

Importance of attributes: Not all attributes for a given concept are equally important, hence, we add features that capture the importance of attributes for a concept. For each attribute–concept pair, Biperpedia includes supporting evidence explaining why they are linked. For instance, the concept COUNTRY has an attribute COFFEE PRODUCTION, and Biperpedia recorded all the countries for which it found mentions of coffee production in text or in queries (e.g., Brazil). The evidence for concept c and attribute a is aggregated using several measures: (1) $instances(a, c)$ is the number of unique supporting instances found for concept c; (2) $frequency(a, c)$ is the number of occurrences of a with supporting instances in the corpus (e.g., the occurrence of Brazil with COFFEE PRODUCTION); and (3) $rank(a, c)$ is the rank of a, ordered by instances and frequency.

For each candidate pair, c_1 and c_2, we consider the set of intersecting attributes I shared by the two concepts. We look at the coverage of attributes in I relative to all attributes of each concept, in terms of their *frequency*, *instances*, and the following *rank* coverage (shown for c_1),

$$\text{RankCoverage}(I) = \sum_{a \in I} \frac{1}{rank(a, c_1)} \quad (1)$$

We distinguish between two sets of intersecting attributes, including all intersecting attributes, or only ones for which subconcept c_1 has fewer or as many supporting *instances*. This distinction addresses transitivity, assuming that, in theory, every instance of the subconcept also supports the superconcept (CMU is a UNIVERSITY and a LOCATION), but the opposite does not hold (LAKE ONTARIO is a LOCATION but not a UNIVERSITY). We compute coverage features based on both intersecting sets.

TF-IDF of attributes: We use an additional adjusted ranking based on the TF-IDF of the frequency of an attribute and concept. TF-IDF is widely used in information retrieval to highlight how important a word is in a specific document, based on a combination of its frequency in the document and its overall uniqueness in the corpus [15]. Here, we consider an attribute a as a word, and a concept c as a document, resulting in a ranking based on:

$$\text{TF-IDF}(a, c) = \log(frequency(a, c) + 1) \cdot \log\left(\frac{|\mathcal{C}|}{|\{c' \in \mathcal{C} : frequency(a, c') > 0\}| + 1}\right) \quad (2)$$

We include coverage features similar to the ones described above using TF-IDF based frequency and ranking.

Attribute synonyms: Biperpedia provides synonyms and common misspellings for each attribute. When comparing the attributes of two concepts, we match attributes using an exact match, and also allow for synonym-based matching. For example, if concept c_1 has the attribute BOAT and concept c_2 has the attribute YACHT and these are detected by Biperpedia as synonyms, we consider them as matching attributes. We replicate the above attribute-based features where new features use synonym matching. In total, we use 97 attribute-based features for each candidate concept pair.

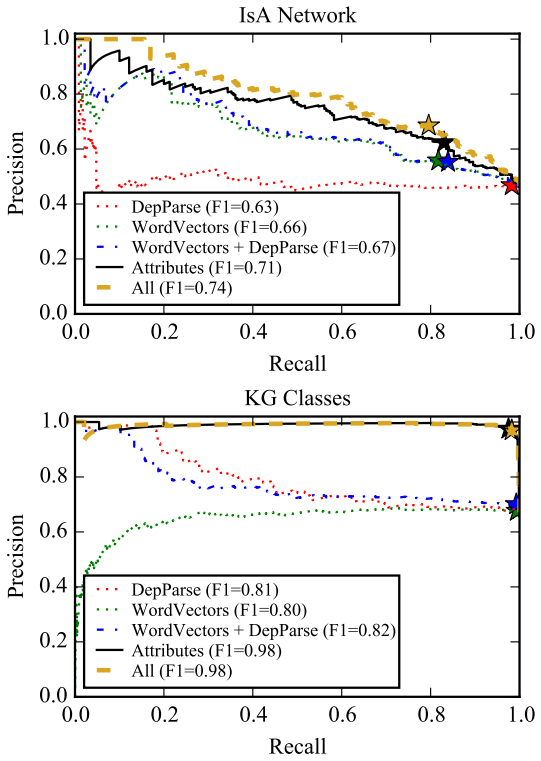

Figure 4: ROC curves of classifiers trained on relations from the IsA Network or KG Classes, using different subsets of features (see legend). A star marks the highest F1, also noted in parentheses.

4. EXPERIMENTAL EVALUATION

We evaluate LATTE on two sets of concepts, one from a large network of automatically extracted IsA relations, and another from a smaller, manually curated knowledge base.

4.1 Latte for the IsA Network

We use a set of 17M concepts extracted from Web text along with 493M *hypernym–hyponym* relationships. We call it an *IsA Network*, as it was extracted using IsA patterns: e.g., "X is a Y", indicates X is a subconcept of Y [25]. This collection contains many concepts, but is inherently noisy.

We applied LATTE to concepts from this set that have at least one known attribute. The IsA Network does not distinguish subsumption from instance-of relations, therefore, we include only concepts with more than 10 hyponyms in the data, which indicates they are not instances, leaving 47K concepts. As the network was generated automatically, we first manually estimated its quality by asking 5 experts to label 5K randomly selected IsA relations from the data. According to the experts, it has 0.54 precision, meaning that while it provides a rich set of concepts, we need to overcome the noise inherent to the extraction process. We found unanimous agreement among three experts on whether an IsA relation is valid in 74% of a random set of 1K relations.

Method	IsA Network	KG Classes
Dep. Parse	0.634	0.811
Word Vectors	0.663	0.805
Dep. Parse & Word Vectors	0.667	0.821
Attributes	0.713	0.975
All	**0.736**	**0.976**
Expert Agreement	0.74	N/A

Table 3: Best F1 results for subsumption classifiers on labeled relations from the IsA Network and KG Classes, using different subsets of features. The last row depicts expert agreement, showing that our classifier achieves accuracy close to expert labeling.

Top-N samples:	100	1K	10K	100K	1M	2.4M
Precision:	1	0.94	0.72	0.68	0.36	0.3

Table 4: Estimated precision at top-N.

To evaluate Latte, we collected 10K subsumption relations based on this IsA network, with 5K extracted directly from the corpus, and 5K created using our candidate extraction process (Section 3.1), all were labeled by 5 experts. We kept all positive examples (2,839) and randomly selected negatives, to give a balanced labeled set. We used a 90%-10% train/test split to evaluate classifiers on labeled data using combinations of the features described above. Figure 4 shows ROC curves on the test set, and the best F1 values are summarized in Table 3. Features based on dependency parsing provide the most limited information. While the semantics of concepts gives strong indication of hierarchy, it lacks coverage, partly since many concept names are short. Dependency parsing is also not fully reliable (e.g., a Hot dog is not a Dog). Surprisingly, word vectors, which reflect the corpus context of a concept, do not perform as well as attribute based context, even when combined with semantic information. Importantly, attributes are the best performing single source of features, with only a minor improvement when combined with other signals. The best F1 of the full system is 0.736, an interesting result given that the agreement of expert labeling on this data was 0.74.

Large-scale prediction analysis: To generate the full set of subsumption relations that Latte can produce on the IsA Network, we use our candidate-generation process on concepts in the full unlabeled network, and predict relations using the best classifier. The result is a concept hierarchy with 2.4M predicted relations. We estimated the accuracy of the hierarchy by manually evaluating 6 sets of 50 random samples from the top predicted relations, by probability (Table 4). Estimated precision on the set of top 100-100K predictions is consistent with our evaluation on the fully labeled test set. While precision on the full set is significantly lower, note that the confidence of this estimate is similarly low due to a small number of samples relative to the set size.

Sample predictions are given in Table 5, including relations predicted in a *directional* way (positively predicted only the input relation), and in a *bidirectional* way (positively predicted the input and reverse relation). Bidirectional prediction indicates we identified synonymy (e.g., ⟨Bass player, Bassist⟩). A sample hierarchy composed of top predicted relations is shown in Figure 1.

Subconcept	Superconcept	p
IsA Network Directional Predictions:		
Mosquito borne disease	Virus	0.993
Blood thinner	Medication	0.993
Fat soluble vitamin	Antioxidant	0.992
IsA Network Bi-Directional Predictions:		
Advertising agency	Ad agency	0.997
Bass player	Bassist	0.997
Malicious software	Malware	0.996
KG Classes Predictions:		
Country albums	Studio albums	0.998
Rugby league teams	Sports teams	0.996
Trade fairs	Conference series	0.996
Newsmagazines	Periodicals	0.995

Table 5: Relations from the IsA Network (top), predicted in a *directional* way or a *bi-directional* way (indicating synonymy), and their probability (p). Bottom rows show predictions from KG Classes.

Category	Example Subconcept-Superconcept	#
Generic/indirect relation	⟨University department, Representation⟩	9
Reverse relation	⟨Health care practitioner, Physician⟩	7
True for few or some instances	⟨Blood borne pathogen, Severe illness⟩	7
Different specialization	⟨Barbiturate, Cholesterol lowering drug⟩	4
Related	⟨Etiology, Inflammatory disorder⟩	19
Unrelated	⟨Opposition party, Purpose ⟩	4

Table 6: Error analysis.

Error analysis: We evaluated a sample of 50 predictions extracted from the 2.4M predicted relations, unanimously labeled as an error. Table 6 shows categories of errors containing at least four examples. Remaining relations were grouped based on whether their concepts were *Related* or *Unrelated*. Note that out of 50 examples, only 4 contained unrelated concepts. The biggest category contains predictions where the superconcept is generic, or the relation is otherwise indirect. These errors were expected based on the level of disagreement among evaluators over generic concepts. Some predictions hold a reverse subsumption relation: a Physician is a Health care practitioner, but we predicted the reverse. Other relations are true for a subset of instances: some Blood borne pathogens (e.g., HIV, Ebola) cause Severe illness, while others are mild. In some cases, the subconcept has a different specialization of the superconcept: Barbiturate is a sedative Drug, however, it does not lower cholesterol.

In many of these errors the predicted concepts are highly-related; it may be that their *distinguishing* aspects are not reflected by their attributes. This can happen for attributes not explicitely mentioned in any Web text (sometimes named *common sense* facts, they are notoriously hard to mine [30]), or if they are discarded due to low occurrence frequency.

4.2 Latte for KG Classes

We use a second set of concepts from a manually curated schema extending Freebase [3]: a structured knowledge base containing millions of facts, organized by 10K types (concepts), with 10K subsumption relations defined

between them. We call this set *KG Classes*.

We applied LATTE to concepts from KG. As positive training examples, we used the manually created relations among concepts in the set, keeping the 5K relations whose concepts appear in our text corpus and have known attributes. We then randomly select 5K examples, likely to be negative, from the following sources: (1) The reverse of the positive examples. (2) An additional Knowledge Graph resource of concepts that are overlapping, but not subsumed. (3) Candidates created with our candidate generation method; while some of these are true subsumptions, any random pair is likely negative. We use 90% of all examples for training, and manually label the 10% used for testing. Attribute-based features (Figure 4 and Table 3) perform markedly better than the baselines, giving an F1 of 0.975, versus 0.976 using all features. This data contains concepts with long names, uncommonly found in a text corpus, causing the word vector method to perform relatively poorly. In comparison, the elaborate semantics of the names results in a better performing dependency-parse classifier. Using the full pipeline we predict all subsumption relations over this data; our predictions constitute a 7% addition to the high-quality relations manually created for this set (samples shown in Table 5).

Comparison of hierarchies: Our results show that starting with higher-quality data (as in KG), we get a higher-quality hierarchy. In comparison, the IsA network is larger and noisier and it produces a larger hierarchy (2.4M relations vs 734 in KG), of reasonably high quality. This difference reflects the tradeoff between diverse and reliable concepts.

5. FINDING HIGH-QUALITY CONCEPTS

We show the ability of LATTE in detecting high-quality concepts from a noisy resource. The IsA network described above is noisy, its relations estimated at 54% precision. Here, we wish to extract from it concepts with high-precision IsA links, showing LATTE can be used to clean up the network.

We select the top 1,000 pairs of concepts identified by LATTE on the network and extract the 50 most frequently occurring concepts in this set. Using solely the IsA network resource, concepts for which IsA relations were extracted with high frequency, should presumably have high quality. Therefore, as a baseline, we extract 50 concepts with comparable IsA relation frequency to the 50 extracted using LATTE. As a second baseline, we consider the 50 most frequent concepts in the network. Our analysis is based on the top 10 hypernyms and hyponyms of each of the 150 concepts.

We manually evaluate the precision of all 3,000 IsA relations; the precision of a concept is the average precision of its IsA relations. The improvement in quality is marked: LATTE identified concepts with higher quality hypernyms, hyponyms, and total IsA relations (Table 7). Figure 5 shows a histogram of average concept quality using all IsA relations. The results indicate that attribute-based relations can be used to increase the quality of existing IsA resources—our pipeline extracted concepts with high IsA relation precision, which could not be detected using frequency.

6. RELATED WORK

Hierarchy construction follows two approaches, which aim at discovering semantic relations between pairs of concepts, forming the backbone of a hierarchy. In the first approach, subsumption relations are extracted based on lexical patterns, correlated with a hierarchical relation (such as "X is a Y") [11], similar to the IsA Network described above. Similar semantic relation classifiers have been introduced for meronyms, synonyms and other analogy relations [9, 12, 27]. Here, we compare LATTE to a similar rule-based method, using the dependency parse tree of concept names.

The main drawback of pattern-based methods, is that they rely on direct evidence appearing in text and lack global context. Conversely, the second approach considers distributional similarity, based on the semantic context of concepts. Hierarchy learning in this case is based on clustering or classification [29, 7, 19, 22, 21], where the context distribution is often evaluated using mutual information, the log-likelihood ratio, cosine similarity or relative entropy [6, 23, 24, 14]. In Section 3.2.2 we explored the recent success of word vector representations, used to account for word similarities. In comparison, we show *attributes* are useful as semantic context, for the task of hierarchy building. LATTE presents a hybrid approach to ontology construction, combining a rule-based signal with distributional similarity based on corpus and attribute context. This combination allows us to overcome the coverage and noise limitations common when using the base methods on data from the Web.

	Latte	**Similar Frequency**	**Most Frequent**
Hyponyms	**0.97** (0.07)	0.55 (0.3)	0.87 (0.2)
Hypernyms	**0.68** (0.23)	0.6 (0.23)	0.54 (0.24)
Total	**0.83** (0.12)	0.58 (0.2)	0.7 (0.18)

Table 7: Average concept quality (with standard deviations) of concepts from LATTE and from the IsA network (similarly or most frequent).

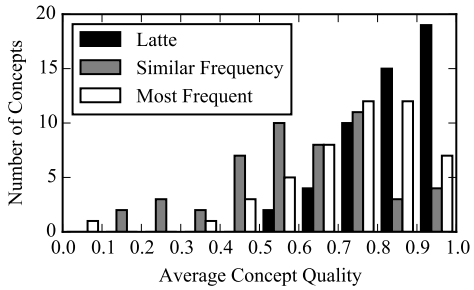

Figure 5: Histogram of average concept quality in LATTE (black), and in the similarly (gray) or most (white) frequent concepts from the IsA network.

7. CONCLUSIONS

We describe a system that finds high-quality subsumption relations among given concepts. These relations try to capture a latent conceptualization of how millions of users query the Web. Our evaluation shows that signals based on concept attributes contribute significantly to detecting high precision relations, out performing state-of-the-art baseline methods. Our system achieves an F1 of 74% on concepts from a network of IsA relations, initially estimated at 54% precision, and an F1 of 98% using higher-quality input concepts. We further illustrate our method finds concepts with high-precision IsA relations, showing that an attribute-based signal can be used to clean resources with noisy concepts.

8. REFERENCES

[1] K. Bellare, P. P. Talukdar, G. Kumaran, F. Pereira, M. Liberman, A. McCallum, and M. Dredze. Lightly-supervised attribute extraction. *NIPS 2007 Workshop on Machine Learning for Web Search*, 2007.

[2] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Information processing & management*, 43(4):866–886, 2007.

[3] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM, 2008.

[4] A. Brown. Get smarter answers from the knowledge graph. `http://insidesearch.blogspot.com/2012/12/get-smarter-answers-from-knowledge\_4.html`, 2012.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[6] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[7] J. R. Curran. *From distributional to semantic similarity*. PhD thesis, University of Edinburgh. College of Science and Engineering. School of Informatics., 2004.

[8] B. Fazzinga, G. Gianforme, G. Gottlob, and T. Lukasiewicz. Semantic web search based on ontological conjunctive queries. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):453–473, 2011.

[9] R. Girju, A. Badulescu, and D. Moldovan. Learning semantic constraints for the automatic discovery of part-whole relations. In *North American Chapter of the Association for Computational Linguistics on Human Language Technology*. ACL, 2003.

[10] R. Gupta, A. Halevy, X. Wang, S. E. Whang, and F. Wu. Biperpedia: An ontology for search applications. *Proceedings of the VLDB Endowment*, 7(7), 2014.

[11] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992.

[12] D. Lin, S. Zhao, L. Qin, and M. Zhou. Identifying synonyms among distributionally similar words. In *IJCAI*, 2003.

[13] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

[14] A. Maedche and S. Staab. Measuring similarity between ontologies. In *Knowledge engineering and knowledge management: Ontologies and the semantic web*, pages 251–263. Springer, 2002.

[15] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[17] T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

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

[20] M. Pasca and B. Van Durme. What you seek is what you get: Extraction of class attributes from query logs. In *IJCAI*, volume 7, pages 2832–2837, 2007.

[21] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *ACL*, 1993.

[22] H. Poon and P. Domingos. Unsupervised ontology induction from text. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, pages 296–305. Association for Computational Linguistics, 2010.

[23] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *arXiv preprint arXiv:1105.5444*, 2011.

[24] P. Senellart and V. D. Blondel. Automatic discovery of similar words. In *Survey of Text Mining*. Citeseer, 2003.

[25] R. Snow, D. Jurafsky, and A. Y. Ng. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*, 2004.

[26] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics, 2011.

[27] P. Turney, M. L. Littman, J. Bigham, and V. Shnayder. Combining independent modules to solve multiple-choice synonym and analogy problems. 2003.

[28] P. D. Turney, P. Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188, 2010.

[29] A. Vinokourov and M. Girolami. A probabilistic framework for the hierarchic organisation and classification of document collections. *Journal of intelligent information systems*, 18(2-3):153–172, 2002.

[30] L. Von Ahn, M. Kedia, and M. Blum. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 75–78. ACM, 2006.

[31] D. C. Wimalasuriya and D. Dou. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 2010.

[32] W. Y. Zou, R. Socher, D. M. Cer, and C. D. Manning. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pages 1393–1398, 2013.