Refer-to-as Relations as Semantic Knowledge

Song Feng
IBM T. J. Watson Research Center &
Stony Brook University
sfeng@us.ibm.com

Sujith Ravi, Ravi Kumar
Google
Mountain View, CA
sravi@google.com ravi.k53@gmail.com

Polina Kuznetsova
Computer Science Department
Stony Brook University
polina.sbu@gmail.com

Wei Liu, Alexander C. Berg, Tamara L. Berg
Computer Science Department
University of North Carolina at Chapel Hill
{wliu, aberg, tlberg}@cs.unc.edu

Yejin Choi
Computer Science & Engineering
University of Washington
yejin@cs.washington.edu

Abstract

We study *Refer-to-as* relations as a new type of semantic knowledge. Compared to the much studied *Is-a* relation, which concerns factual taxonomic knowledge, *Refer-to-as* relations aim to address pragmatic semantic knowledge. For example, a "penguin" *is a* "bird" from a taxonomic point of view, but people rarely *refer to* a "penguin" as a "bird" in vernacular use. This observation closely relates to the *entry-level categorization* studied in Psychology. We posit that Refer-to-as relations can be learned from data, and that both textual and visual information would be helpful in inferring the relations. By integrating existing lexical structure knowledge with language statistics and visual similarities, we formulate a collective inference approach to map all object names in an encyclopedia to commonly used names for each object. Our contributions include a new labeled data set, the collective inference and optimization approach, and the computed mappings and similarities.

Introduction

We study *Refer-to-as* relations as a new type of semantic knowledge. We define *Refer-to-as(A,B)* to denote the pragmatic knowledge about the language use such that an object A is typically referred to as B in vernacular use. Compared to the much studied *Is-a* relations (Hearst 1992), which concern strictly factual knowledge of taxonomy, *Refer-to-as* relations aim to address pragmatic semantic knowledge that is crucial for practical language understanding and production systems. For example, a "cat" is a "carnivore" from a taxonomy point of view (*Is-a(cat, carnivore)*), but people rarely refer to a "cat" as a "carnivore" in most conversational settings (¬*Refer-to-as(cat, carnivore)*).

This closely relates to *entry-level categories* originally introduced in pioneering work from psychologists including Eleanor Rosch (Rosch 1978) and Stephen Kosslyn (Jolicoeur, Gluck, and Kosslyn 1984). Entry-level categories are the names people tend to associate with objects in the world.

We posit that the entry-level categorization can drive a new type of semantic knowledge, *Refer-to-as* relations. Having this knowledge at scale could provide a useful enhancement to existing linguistic resources, like WordNet (Fellbaum 1998), by adding *Refer-to-as* relations that complement existing *Is-a* (hypernym) and *Has-a* (meronym) relations. Automatically extracted *Refer-to-as* relations could also be useful in a number of natural language applications ranging from coreference resolution of nominal nouns, to referring expression generation in human robot interactions, to abstractive summarization or text simplification with lexical substitutions.

Experiments from psychology (Rosch 1978; Jolicoeur, Gluck, and Kosslyn 1984) in the 70s and 80s have provided insights into this natural cognitive conceptualization, which are based on questions such as "what would you call this object?" or "is this object a kind of X?". However, due to the nature of human-subject-based psycho-physical experiments, these studies were relatively small in scale, considering only on the order of tens of object categories.

Computational methods for modeling entry-level categorization remain largely unstudied. A notable exception is the recent work of (Ordonez et al. 2013) who explore initial methods to translate between encyclopedic and entry-level categories using visual models and/or linguistic resources. We take this idea much further in several dimensions. First, we propose *Refer-to-as* relations to be studied as a new conceptual problem in semantic knowledge, with the distinct purpose of learning the relations over *all* object names in an encyclopedia knowledge base. Second, we posit that natural cognitive conceptualization is inherently a collective inference process, and propose a formulation that infers *Refer-to-as* relations over all encyclopedia categories jointly. Additionally, we make use of recent deep learning based visual features (Krizhevsky, Sutskever, and Hinton 2012; Girshick et al. 2013) to measure the visual similarity of concepts.

Building on (Ordonez et al. 2013) we present a combinatorial optimization formulation to jointly infer *Refer-to-as* relations. The objective function encodes some of the key aspects of entry-level conceptualization, in particular: (1) *generality* (quantified via *frequency*) of a category name, and (2) *representativeness* (quantified via *semantic proximity* and/or *visual similarity*) between an encyclopedic category and a putative entry-level concept. For the optimization problem, we introduce solutions based on integer linear programming and minimum-cost flow. Experimental results show that the proposed approaches outperform competitive baselines, confirming our hypothesis to view entry-level categorization as a collective inference problem over

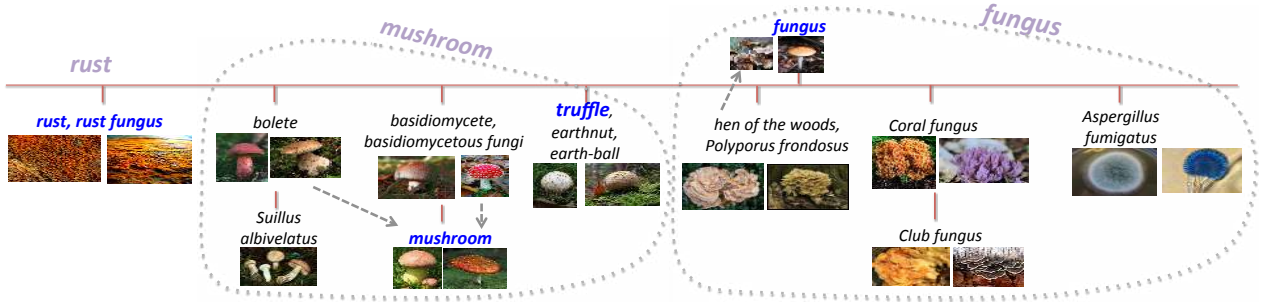

Figure 1: WordNet hierarchy and entry-level categories (in bold blue smaller font).

lexical structure knowledge. We are making the resulting entry-level categories and visual similarities publicly available at http://homes.cs.washington.edu/∼yejin/refer2as/

Insights and Definitions

Insights into an Existing Taxonomy We use WordNet (Fellbaum 1998) as the encyclopedia knowledge base. WordNet taxonomy, as defined by lexicographers, concerns strictly factual knowledge, which does not always align with commonsense or pragmatic knowledge. Figure 1 shows a simplified hierarchical structure under the branch of "fungus". Per WordNet, "truffle" is a "fungus", but not a "mushroom", because "mushroom" only appears as a descendant of "basidiomycete", which is a sibling of "truffle". By studying the task of learning *Refer-to-as* relations, we aim to complement the existing taxonomies such as WordNet with pragmatic semantic knowledge.

Encyclopedia Categories Prototype Theory in literature has been developed primarily in the context of physical categories, as the notion of prototypes for abstract categories (e.g., semantics, derivation) is not as obvious. We also focus our study on the physical objects and consider all inherited hyponyms of the *physical_entity.n.01* synset in WordNet 3.0, which amounts to 39,556 categories in total.

Due to the innate structure of WordNet, each encyclopedia category at this point is mapped to a *synset*, which, in turn, is defined as a set of words. For practical convenience, we represent each encyclopedia category by the first word in the synset, which generally corresponds to the word that is used the most often for the specific sense. Conversely, in what follows, when we need to map a word to a synset, we choose the first synset listed under the given word.

Entry-Level Categories These correspond to a subset of the physical entity categories that are more commonly used as surrogates of the specific ones (encyclopedia categories). We assume that each encyclopedia category corresponds to at least one entry-level category, which could be the encyclopedia category itself.

Modeling Entry-Level Categorization

Task Definition The goal in entry-level categorization is the following: Given a set $\mathcal{C}_{\text{encyc}}$ of encyclopedia categories and a list $\mathcal{C}^i_{\text{entry}}$ of entry-level candidates for each category $i \in \mathcal{C}_{\text{encyc}}$, the task is to find an entry-level category assignment $\sigma(i) \in \mathcal{C}^i_{\text{entry}}$ for all encyclopedia categories in $\mathcal{C}_{\text{encyc}}$.

Quantifying Generality One of the key aspects of entry-level categorization is the notion of *generality* of the concept. We use word frequencies derived from Google Ngram 1T data (Brants and Franz. 2006), where the generality of a category i is quantified as the normalized log frequency $\text{freq}(i)$ scaled to be $\in$ [0,1].

Path-based Representativeness An entry-level category should be conceptually *representative* for its encompassing encyclopedia categories. As an approximate measure, we use the LCH score (Leacock, Miller, and Chodorow 1998), which quantifies the semantic proximity of two concepts (senses) by the shortest path between them normalized by the depth of those concepts in the taxonomy.

Visual Representativeness We compute visual similarity as another measure of *representativeness*, using the ImageNet dataset (Deng et al. 2009) which contains images illustrating 21,841 of the WordNet synsets. We build the notion of synset similarity on top of an image similarity function $f : (I, J) \mapsto r \in [-1, 1]$ that compares pairs of images. In order to deal with variation in appearance between object instances we define the similarity between image sets $\mathcal{I}$ and $\mathcal{J}$ as: $\frac{1}{2}\left(d\left(\mathcal{I},\mathcal{J}\right) + d\left(\mathcal{J},\mathcal{I}\right)\right)$, where $d\left(\mathcal{J},\mathcal{I}\right) = \frac{1}{5|\mathcal{I}|}\sum_{I\in\mathcal{I}}\sum_{k=\{1,\ldots,5\}} f(I, J^I_k)$ and J^I_k indicates the image J in $\mathcal{J}$ with the kth largest value for $f(I, J)$.

To compute the similarity between a pair of images, (I, J), we use the open source Caffe (Jia 2013) software that implements the groundbreaking convolutional neural network (convnet) architecture of (Krizhevsky, Sutskever, and Hinton 2012). In particular, to process an image I, we resize to 256×256, pass it through the pre-trained Caffe model, and extract the 6th layer response of the model, then (L2) normalize to produce the 4096 dimensional feature descriptor $d(I)$. Then $f(I, J) = \langle d(I), d(J)\rangle$. We download and process all (21841) synsets from the ImageNet Fall 2011 release, sampling 100 images per synset. This involves computing Caffe features for over 2 million images and computing their pairwise similarities.

Candidate Entry-Level Selection We obtain a set $\mathcal{C}^i_{\text{entry}}$ of candidate entry-level categories for an encyclopedia category i as follows:

(i) Head ($\mathcal{C}^{\text{head}}$): For an encyclopedia category whose name corresponds to a multi-word phrase, we include the head word whose frequency $\geq \theta^{\text{freq}}$ as a candidate for the entry-level category. We set θ^{freq} to 0.3 based on the distribution of frequencies of entry-level categories in the development dataset.

(ii) Self ($\mathcal{C}^{\text{self}}$): If an encyclopedia category has a generality score $\geq \theta^{\text{freq}}$, we consider the category itself as an entry-level category candidate.

(iii) Hypernyms ($\mathcal{C}^{\text{hyper}}$): Since WordNet hierarchy provides a good starting point for candidates, we include five inherited hypernyms with the highest generality scores such that $\#hyponyms \leq \theta^{\text{hypo}} = 5000$.

(iv) Relatedness score ($\mathcal{C}^{\text{rel}}$): We also include top ten categories with the highest LCH scores such that $\#hyponyms \leq \theta^{\text{hypo}}$. This allows us to include words outside the hypernym path as candidates, mitigating the idiosyncrasy of the WordNet taxonomy as discussed earlier.

Combinatorial Formulation

In the entry-level categorization problem, the goal is: for each encyclopedia category, assign an entry-level category from its candidate list. In the assignment, we wish to simultaneously maximize two objectives:

(i) Representativeness: The chosen entry-level categories should be *representative* of the encyclopedia categories to which they are assigned; we model this using a semantic similarity matrix.

(ii) Generality: The set of entry-level synsets should capture prototypes that are *commonly* used by humans to refer to objects; we use a popularity measure for this purpose.

Given this, we formulate the entry-level categorization task as a constraint optimization problem.

Formulation Let A be an $m \times n$ non-negative matrix and let B be a non-negative vector of size n. We assume that all the entries of A and B are in $[0, 1]$. The goal is to find an assignment $\sigma : [m] \to [n]$ to maximize:

$$\sum_{i=1}^{m} A_{i,\sigma(i)} + \sum_{j=1}^{n} f\left(B_j, |\sigma^{-1}(j)|\right), \tag{1}$$

where f is a fixed function. The above formulation can be applied to a broad category of problems. Specifically, for the entry-level categorization task, we model the semantic representativeness and generality using A and B respectively and carefully choose f. Note that the second component in the objective measures something global, e.g., in our task, depending on f, this can capture the overall popularity of individual entry-level candidates. Without this second component, problem (1) becomes trivial since the best assignment would be to set $\sigma(i) = \arg\max_j A_{i,j}$.

The integer programming (ILP) version of (1) is:

$$\text{maximize} \sum_{i,j} A_{ij} \cdot p_{ij} + \sum_{j} f\left(B_j, \sum_{i} p_{ij}\right)$$
$$\text{subject to} \sum_{j} p_{ij} = 1, \forall i; \quad p_{ij} \in \{0, 1\}, \forall i, j.$$

This ILP problem can be solved using highly optimized solvers (CPLEX 2006) and the solution will yield an entry-level assignment for our original problem. (See (Schrijver 1986) for more details on ILP.) Unfortunately, ILPs can be expensive to solve on large problem instances. Hence, we also consider algorithms that are computationally more efficient. (Notice that it is trivial to obtain a 2-approximation to (1) by considering the best solution to either parts of the objective; the first component by itself can be solved exactly and for f's we will consider, we can solve the second component exactly as well.)

One reasonable candidate for f is

$$f(b, s) = b \cdot [s > 0], \tag{2}$$

i.e., the second component in (1) becomes $\sum_{j \in \text{range}(\sigma)} B_j$. Unfortunately, with this f the problem becomes computationally hard.

Lemma *Objective (1) with (2) is NP-hard.*

Proof. Consider the exact cover by 3-sets (X3C) problem: given a universe U of m elements and a collection $\mathcal{F}$ of n subsets, each of size three, cover all the elements of U by a subset of $\mathcal{F}$ such that each element is covered exactly once; this problem is NP-complete (see, for example, (Papadimitriou 1994)). Let $k = \max(m, n) + 1$. Given an instance of X3C, construct the $m \times n$ incidence matrix A where $A_{u,F} = k$ if and only if $u \in F$ and let $b(F) = 1$ for all $F \in \mathcal{F}$. it is easy to see that the solution to (1) is $mk + m/3$ if and only if the given X3C instance has a solution. Indeed, if the X3C instance has a solution $\mathcal{F}^* \subseteq \mathcal{F}$, then set the assignment to be $\sigma(u) = F$ where $u \in F$ and $F \in \mathcal{F}^*$; since it is an exact cover (i.e., $|\mathcal{F}^*| = m/3$), the value of this solution is $mk + m/3$. Conversely, by the choice of k, any assignment σ to (1) of value $mk + m/3$ will have the property that $A_{i,\sigma(i)}$ has to be non-zero (in fact, equal to k) for every i. Thus, the first term of (1) will incur a contribution of mk. Now, since $B(F) = 1$, we have $|\text{range}(\sigma)| = m/3$, which means an exact cover.

If instead we make the contribution of an entry-level category *proportional* to the number of encyclopedia categories to which it can be assigned,

$$f(b, s) = g(b) \cdot s, \tag{3}$$

for some g, then the problem becomes min-cost flow.

Min-Cost Flow Formulation Recall that in the min-cost flow problem, we are given a graph where the edges have costs and capacities and a desired flow value. The goal is to achieve this flow with the minimum cost subject to the capacity constraints (Ahuja, Magnanti, and Orlin 1993).

We construct the following min-cost flow instance from A and B. The main idea is to use the value B_j to define the cost and capacity constraints for $j \in [n]$. Specifically, we consider the bipartite graph implied by A; the capacity of all the edges in this bipartite graph is 1 and the cost is $1 - A_{ij}$. We add a source node s and connect to all nodes in $[m]$ with cost 0 and capacity 1. We add a sink node t and connect it to all nodes in $j \in [n]$ with cost $1 - g(B_j)$ and capacity $\deg(j)$, which is the number of non-zero entries in the jth column of A. Since we require an assignment, we solve for

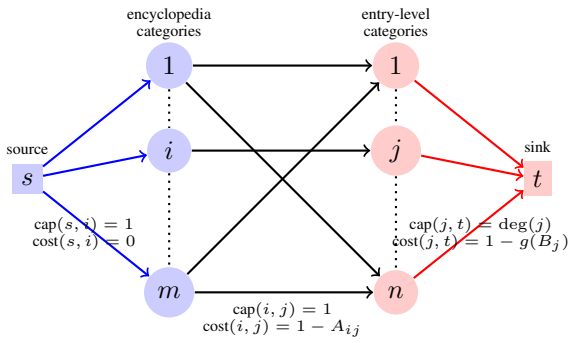

Figure 2: Entry-level categorization as an (s,t)-min-cost flow problem with capacities and costs.

the minimum cost flow from s to t with value m (Figure 2). It can be seen that the min-cost flow solves (1) with (3).

The advantage of a min-cost formulation is that it can be solved in polynomial time. In our experiments, min-cost flow took under a minute while yielding comparable assignment quality whereas CPLEX took an order of magnitude more time to solve an ILP instance for the same dataset.

We now comment on the choice of the function g. An obvious first choice is $g(b) = b$. A much better choice is to carefully model the contribution of *rare* versus *more* popular candidates with a sigmoid-style function given by

$$g(b) = 1 - \sqrt[k]{\frac{\theta - \min(b,\theta)}{a}},$$

where θ, a, k are parameters. The idea is to reward popular category assignments while attenuating the contribution of less popular entry candidates ($b < \theta$) as a non-linear function of their popularity. We use $\theta = 0.5, a = 2, k = 2$. This formulation of max-flow assumes one *Refer-to-as* relation for each encyclopedic category. We leave it as future work to relax this assumption to accommodate a set of relations.

Evaluation Design

Because entry-level categorization has not been studied before as a lexical semantic knowledge problem, we develop a suitable dataset for evaluation. It is worthwhile to discuss the labeled dataset of (Ordonez et al. 2013) first. As their study was driven by *visual recognition*, rather than *semantic knowledge*, a natural design of human annotation was to show sample images of an ImageNet node, and ask *what are present* in those images. However, due to the inherent ambiguities in what these images contain and represent, care should be taken in interpreting the resulting labels as *Refer-to-as* relations. For example, for the ImageNet node "volleyball net", half of the images contain volleyball players, while others look like desolate beach scenes, misleading turkers to identify "people" or "beach" as the entry-level categories of "volleyball net". Removing this ambiguity is important in our study, as we intend to draw more explicit connections between labeled entry-level names to a specific encyclopedia concept as *Refer-to-as* relations. For this reason, in all our human annotation studies, we present sample images together with the textual descriptions (gloss) for each node.

Using Amazon Mechanical Turk (AMT), we collect two different types of human annotations: (I) Fill-in-the-blank and (II) Multiple-choice:

LabelSet-I: Fill-in-the-blank For each encyclopedic object name, we ask turkers to think of commonly used names (i.e., how people would generally refer to the given object in real life scenarios). Considering that turkers may be unfamiliar with some of the encyclopedic object names, e.g., "basidiomycete", we provide the following set of information to help defining the concept: (1) the name of the encyclopedic category; (2) the definition of the encyclopedic object obtained from WordNet (Fellbaum 1998), and (3) five corresponding images obtained from ImageNet (Deng et al. 2009). For each encyclopedic category, we ask five turkers to provide up to three labels. This results in an average of 3.5 unique names for each category.

LabelSet-II: Multiple-choice The potential problem of *fill-in-the-blank* type labels is that it is hard to collect a complete set of valid names. Therefore, when a system identifies an entry-level name not specified by a turker, we cannot be certain whether it is an invalid choice or a good choice that turkers forgot to include. We therefore also include *multiple-choice* annotation. The set of choices includes any potential name that any of the systems can output for a given encyclopedia category. For each category, we ask three turkers to choose one of the following: (1) "Yes, I would use this term", (2) "Possibly yes, though I would use a more *general* (or *specific*, given as a separate option) term", (3) "No", and (4) "I don't know".

Baseline-QA This baseline helps us gain empirical insights into how solutions to *Is-A* relations perform on a related, but very different task of learning *Refer-to-as* relations. Based on the work of (Prager, Radev, and Czuba 2001) for "what is" question-answering in the context of keyword-based search, we first obtain the co-occurrence information of a term and its hypernyms from Google Ngram 1T corpus. $\sim$12,000 out of $\sim$40,000 encyclopedia terms have non-zero co-occurrence with at least one hypernym. For those cases, we calculate a score by normalizing the co-occurrence by the "level number" as described in (Prager, Radev, and Czuba 2001), and then select the hypernym with highest score as entry-level category (the answer of "what is" question).

Baseline-QA-Self The baseline described above performs poorly in large part because it does not allow "self" as the potential answer to the query term. (There is no co-occurrence statistics for "self".) Therefore, we augment the above baseline to assign the term with relatively high Google frequency ($\text{freq}(i) \geq 0.5$) as the entry-level category for itself (BASELINE-QA-SELF).

Upper Bound There are 90.1% of encyclopedia categories whose candidates match at least one entry-level names in LabelSet-I while the rest matches none.

METHODOLOGY	PRECISION
BASELINE-QA	56.14
BASELINE-QA-SELF	69.49
ORDONEZ ET AL.[3]	64.07
BASELINE	65.02
FLOW(τ)	69.29
FLOW(ν)	**71.03**
FLOW(τ, ν)	**71.16**
ILP(τ)	61.51
ILP(ν)	64.19
ILP(τ, ν)	62.66
ILP(τ)^C	70.40
ILP(ν)^C	69.31
ILP(τ, ν)^C	69.76

Table 1: Evaluation with LabelSet-I.[1]

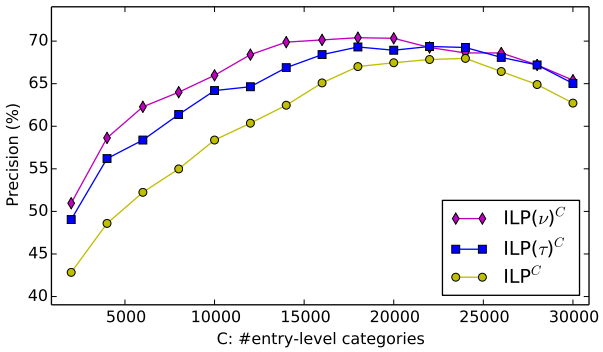

Figure 3: Accuracy vs. #entry-level categories.

Experimental Results

We experiment with varying composition of $A_{i,j}$ in (1) as:

$$A_{i,j} = \tau \cdot S_{i,j}^{\text{textual}} + \nu \cdot S_{i,j}^{\text{visual}} + K_j$$

where $\tau, \nu \in \{0, 0.5, 1\}$, and $K_j \leftarrow 1$ if $j \in \mathcal{C}^{\text{head}}$.

Evaluation with LabelSet-I Table 1 shows the cross-validated performance based on LabelSet-I. (τ) and (ν) indicate that the terms corresponding to the textual and visual similarities are activated respectively. Flow algorithms achieve the best performance, with a substantial margin over the baseline approaches.

How Many Entry-Level Names for WordNet? The ILP formulation allows us to set a hard constraint C on the total number of entry-level categories selected by the solver. As shown in Table 1, this constraint improves the performance significantly over counterparts. Figure 3 shows a common trend across various configurations of $A_{i,j}$ such that the peak performances are achieved for $C \in [19000, 22000]$. For flow algorithms, we do not have the cardinality constraint. Interestingly, the best performing flow variant also selects about 21,000 as entry-level categories. This indicates that there might be a natural range of cardinality for entry-level categories for all encyclopedia names in WordNet.

[1]Note that the comparison against (Ordonez et al. 2013) is not apple-to-apple in that their methods are not tuned for our task and evaluation.

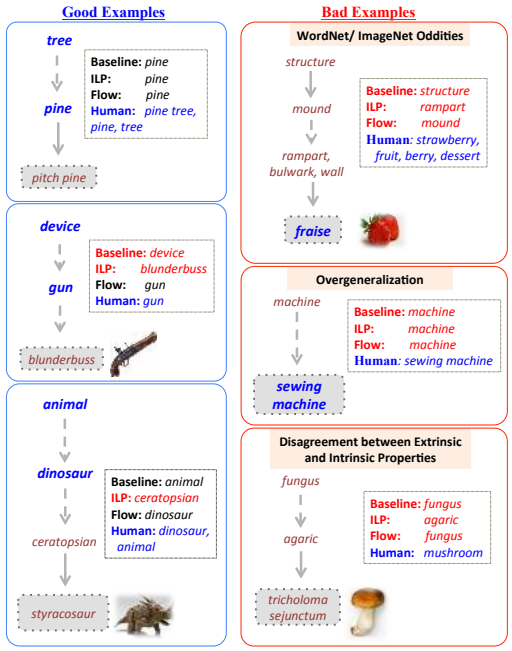

Figure 4: Examples of system output.

FLOW(τ) > FLOW (ν)	FLOW(ν) > FLOW(τ)
shrub.n.01	substance.n.07
bird.n.01	food.n.01
vehicle.n.01	instrument.n.01
consumer_goods.n.01	solid.n.01
commodity.n.01	mammal.n.01
woody_plant.n.01	container.n.01
fish.n.01	structure.n.01
aquatic_vertebrate.n.01	plant_organ.n.01
invertebrate.n.01	plant_part.n.01

Table 2: The subsets of categories with greatest difference in performance for FLOW algorithm.

Visual Similarities as Lexical Similarities: We further examine the contribution of textual and visual semantic similarities for entry-level categorization over different regions of the WordNet hierarchy. As shown in Table 2, in some categories, visual similarities yield better results than textual similarities, and vice versa. These results suggest the potential utility of visual similarities as a surrogate to quantifying lexical similarities.

Evaluation with LabelSet-II Table 3 shows the evaluation based on LabelSet-II.[2] The numbers under HUMAN shows the evaluation of human labels from LabelSet-I against LabelSet-II configuration, which can be considered as a reference upper bound.

We report the results based on all cases (ALL) and after discarding those cases (MAJORITY AGREES) in which none of the turkers agrees with each other, which corresponds to 26% cases.

[2]LabelSet-II has been collected one time, after we finalized all the algorithmic development.

METHODOLOGY	ALL					MAJORITY AGREES				
	∪ YES	YES	YES, *but should be more* GENERAL	SPECIFIC	NO	∪ YES	YES	YES, *but should be more* GENERAL	SPECIFIC	NO
BASELINE-QA	83.43	43.46	19.56	20.41	16.58	84.85	57.74	12.72	14.39	15.15
BASELINE-QA-SELF	88.44	52.73	19.38	16.33	11.56	91.18	71.14	11.98	8.06	8.82
HUMAN	**92.17**	**57.17**	20.65	14.35	7.82	**94.38**	**75.57**	13.03	5.78	5.61
FLOW(τ, κ)	**91.35**	**54.83**	21.34	15.18	8.65	**93.87**	73.26	13.48	7.13	6.13
FLOW(ν, κ)	90.86	54.02	22.1	14.74	9.14	93.62	72.28	14.5	6.84	6.39
ILP(τ, κ) ^{c}	90.5	54.58	21.42	14.5	9.51	93.59	**73.97**	13.43	6.19	6.41
ILP(ν, κ)^c	89.1	52.9	21.45	14.75	10.9	91.65	70.79	13.87	6.99	8.34
Ordonez et al.	88.88	52.18	20.97	15.73	11.12	90.39	71.26	11.65	7.48	9.62

Table 3: Evaluation with LabelSet-II.

"∪YES" denotes the sum over all variants of YES. For this collective yes (∪YES), the accuracy reaches to 93.87%, close to human performance, achieved by FLOW(τ). For clean yes,[3] the accuracy reaches to 73.97%, which again is close to 75.57% of HUMAN.

Discussion with Examples Examples of the system output are given in Figure 4. While ILP performed reasonably well we can still see a few examples where ILP made a mistake (such as "blunderbuss" or "ceratopsian"), while the min-cost flow guessed the entry-level categories correctly ("gun", "dinosaur"). However, there were a number of cases, when both methods made mistakes. Some errors were due to strange cases in WordNet and ImageNet hierarchy. E.g., "fraise", which is also a French word for "strawberry" in English is defined as "sloping or horizontal rampart of pointed stakes". For this synset, however ImageNet provides pictures of strawberries, even its definition in WordNet is according to English sense of the word. Another interesting example of an erroneous output is "fungus" vs. "mushroom". As we can see from Figure 1 some entities of fungus synset look very much like mushroom, which results in human label "mushroom" for these entities. In a way, there is a disagreement between intrinsic properties of such an entity (its classification in WordNet) and its extrinsic features (visual similarity to mushroom).

Related Work

Cognitive Conceptualization of Categories Although *naming* is an important aspect of how people communicate in natural language and has been studied extensively in Psychology (Rosch 1973; 1978; Lakoff 1987; Taylor 2003), there is little large-scale computational work on finding entry-level categories. Our work is preceded only by a recent work of (Ordonez et al. 2013), but the overall goals, problem formulations, and end results are different. Our end goal is to find natural names for all object names in WordNet and perform naming as a collective inference over all words simultaneously by recognizing the combinatorial nature in the naming assignment. In addition, we explore the new perspective of viewing entry-level categories as a type of lexical semantics and explore the viability of learning entry-level categorization based only on dictionary knowledge and language statistics. In contrast, their work was motivated from the perspective of reusing various visual recognizers and improving object recognition of a given query image. Therefore, instead of targeting all object names in WordNet, they mainly operated with a subset of nodes for which reliable visual recognizers are available.

Categorization in Language Acquisition Another line of research that touches on the notion of cognitive conceptualization of natural language words is that of child language acquisition (Saffran, Senghas, and Trueswell 2001). The main difference is that we aim to learn how humans name the extensive list of objects in the world, while conceptualization in early language acquisition focuses on much different linguistic aspects such as syntactic categorization and clustering of similar words with respect to a very limited vocabulary of a child.

Hypernyms The task of learning entry-level categories has a conceptual commonality with that of learning hypernyms (*Is-a* relation). For the latter, there has been a great deal of previous work (Snow, Jurafsky, and Ng 2004).

The key difference is that Is-a relations are "exhaustive" and "taxonomic" relations. For example, for any given leaf-level word in WordNet, there are more than a dozen different hypernyms. In contrast, Refer-to-as relations are much more "selective" and "pragmatic" relations. Selecting one of these spuriously many hypernyms appears to be repeated efforts across many downstream NLP tasks.

Conclusions

In this paper we presented the first computational linguistic study that explores *Refer-to-as* relations as a lexical semantics problem. We first explore constraint optimization based on Integer Linear Programming, which can be computationally expensive. We then present a min-cost flow algorithm that attains as good or better results while solving the optimization exactly and in polynomial time. The key aspect of both these approaches is that they assign *Refer-to-as* relations over all object names collectively while incorporating insights discussed in Prototype Theory in Psychology. Our study confirms the viability of computational and data-driven entry-level categorization, and encourages additional research along this line.

Acknowledgement

We thank reviewers for many helpful comments. Tamara L. Berg is supported by NSF award #1444234 and #1445409.

[3] We assign an entry-level category with a clean yes label if any turker assigned this label.

References

Ahuja, R. K.; Magnanti, T. L.; and Orlin, J. B. 1993. *Network Flows: Theory, Algorithms, and Applications*. Englewood Cliffs, NJ: Prentice Hall.

Brants, T., and Franz., A. 2006. Web 1T 5-gram version 1. In *Linguistic Data Consortium*.

CPLEX. 2006. High-performance software for mathematical programming and optimization. See `http://www.ilog.com/products/cplex/`.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.

Girshick, R. B.; Donahue, J.; Darrell, T.; and Malik, J. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation (v3). *CoRR* abs/1311.2524.

Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, 539–545.

Jia, Y. 2013. Caffe: An open source convolutional architecture for fast feature embedding. `http://caffe.berkeleyvision.org/`.

Jolicoeur, P.; Gluck, M. A.; and Kosslyn, S. M. 1984. Pictures and names: making the connection. cognitive psychology. *Cognitive Psychology* 16:243–275.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Lakoff, G. 1987. *Women, Fire and Dangerous Things: What Categories Reveal About the Mind*. Chicago: University of Chicago Press.

Leacock, C.; Miller, G. A.; and Chodorow, M. 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics* 24(1):147–165.

Ordonez, V.; Deng, J.; Choi, Y.; Berg, A. C.; and Berg, T. L. 2013. From large scale image categorization to entry-level categories. In *ICCV*, 2768–2775.

Papadimitriou, C. M. 1994. *Computational complexity*. Reading, Massachusetts: Addison-Wesley.

Prager, J.; Radev, D.; and Czuba, K. 2001. Answering what-is questions by virtual annotation. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, 1–5. Stroudsburg, PA, USA: Association for Computational Linguistics.

Rosch, E. 1973. Natural categories. *Cognitive Psychology* 4:328–350.

Rosch, E. 1978. Principles of categorization. In Rosch, E., and Lloyd, B., eds., *Cognition and Categorization*. Hillsdale, NJ: Erlbaum. 27–48.

Saffran, J. R.; Senghas, A.; and Trueswell, J. C. 2001. The acquisition of language by children. *PNAS* 98:12874–12875.

Schrijver, A. 1986. *Theory of Linear and Integer Programming*. New York, NY: John Wiley & Sons, Inc.

Snow, R.; Jurafsky, D.; and Ng, A. Y. 2004. Learning syntactic patterns for automatic hypernym discovery. In *NIPS*.

Taylor, J. R. 2003. *Linguistic Categorization*. Oxford Linguistics. OUP Oxford.