In *Proceedings of EMNLP 2009*

Toward Completeness in Concept Extraction and Classification

Eduard Hovy and Zornitsa Kozareva
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
hovy@isi.edu, zkozareva@gmail.com

Ellen Riloff
School of Computing
University of Utah
Salt Lake City, UT 84112
riloff@cs.utah.edu

Abstract

Many algorithms extract terms from text together with some kind of taxonomic classification (is-a) link. However, the general approaches used today, and specifically the methods of evaluating results, exhibit serious shortcomings. Harvesting without focusing on a specific conceptual area may deliver large numbers of terms, but they are scattered over an immense concept space, making Recall judgments impossible. Regarding Precision, simply judging the correctness of terms and their individual classification links may provide high scores, but this doesn't help with the eventual assembly of terms into a single coherent taxonomy. Furthermore, since there is no correct and complete gold standard to measure against, most work invents some ad hoc evaluation measure. We present an algorithm that is more precise and complete than previous ones for identifying from web text just those concepts 'below' a given seed term. Comparing the results to WordNet, we find that the algorithm misses terms, but also that it learns many new terms not in WordNet, and that it classifies them in ways acceptable to humans but different from WordNet.

1 Collecting Information with Care

Over the past few years, many algorithms have been published on automatically harvesting terms and their conceptual types from the web and/or other large corpora (Etzioni et al., 2005; Pasca, 2007; Banko et al., 2007; Yi and Niblack, 2005; Snow et al., 2005). But several basic problems limit the eventual utility of the results.

First, there is no standard collection of facts against which results can be measured. As we show in this paper, WordNet (Fellbaum, 1998), the most obvious contender because of its size and popularity, is deficient in various ways: it is neither complete nor is its taxonomic structure inarguably perfect. As a result, alternative ad hoc measures are invented that are not comparable. Second, simply harvesting facts about an entity without regard to its actual subsequent organization inflates Recall and Precision evaluation scores: while it is correct that a *jaguar* is a *animal, mammal, toy, sports-team, car-make*, and *operating-system*, this information doesn't help to create a taxonomy that, for example, places *mammal* and *animal* closer to one another than to some of the others. ((Snow et al., 2005) is an exception to this.) As a result, this work may give a misleading sense of progress. Third, entities are of different formal types, and their taxonomic treatment is consequently different: some are at the level of instances (e.g., *Michelangelo was a painter*) and some at the level of concepts (e.g., *a painter is a human*).

The goal of our research is to learn terms for entities (objects) and their taxonomic organization simultaneously, from text. Our method is to use a single surface-level pattern with several open positions. Filling them in different ways harvests different kinds of information, and/or confirms this information. We evaluate in two ways: against WordNet, since that is a commonly available and popular resource, and also by asking humans to judge the results since WordNet is neither complete nor exhaustively taxonomized.

In this paper, we describe experiments with two rich and common portions of an entity taxonomy: Animals and People. The claim of this paper is: *It is possible to learn terms automatically to populate a targeted portion of a taxonomy (such as below An-*

imals or People) both at high precision compared to WordNet and including additional correct ones as well. We would like to also report on Recall relative to WordNet, but given the problems described in Section 4, this turns out to be much harder than would seem.

First, we need to define some basic terminology:
term: An English word (for our current purposes, a noun or a proper name).
seed term: A word we use to initiate the algorithm.
concept: An item in the classification taxonomy we are building. A concept may correspond to several terms (singular form, plural form, the term's synonyms, etc.).
root concept: A concept at a fairly general (high) level in the taxonomy, to which many others are eventually learned to be subtypes/instances of.
basic-level concept: A concept at the 'basic level', corresponding approximately to the Basic Level categories defined in Prototype Theory in Psychology (Rosch, 1978). For our purposes, a concept corresponding to the (proto)typical level of generality of its type; that is, a *dog*, not a *mammal* or a *dachshund*; a *singer*, not a *human* or an *opera diva*.
instance: An item in the classification taxonomy that is more specific than a concept; only one example of the instance exists in 'the real world' at any time. For example, *Michelangelo* is an instance, as well as *Mazda Miata with license plate 3HCY687*, while *Mazda Miata* is not.
classification link: We use a single relation, that, depending on its arguments, is either *is a type of* (when both arguments are concepts), or *is an instance of* or *is an example of* (when the first argument is an instance/example of the second).

Section 2 describes our method for harvesting; Section 3 discusses related work; and Section 4 describes the experiments and the results.

2 Term and Relation Extraction using the Doubly-Anchored Pattern

Our goal is to develop a technique that automatically 'fills in' the concept space in the taxonomy below any root concept, by harvesting terms through repeated web queries. We perform this in two alternating stages.

Stage 1: Basic-level/Instance concept collection: We use the Doubly-Anchored Pattern DAP developed in (Kozareva et al., 2008):

DAP: *[SeedTerm1] such as [SeedTerm2] and <X>*

which learns a list of basic-level concepts or instances (depending on whether SeedTerm2 expresses a basic-level concept or an instance).[1] DAP is very reliable because it is instantiated with examples at both 'ends' of the space to be filled (the higher-level (root) concept SeedTerm1 and a basic-level term or instance (SeedTerm2)), which mutually disambiguate each other. For example, "presidents" for SeedTerm1 can refer to the leader of a country, corporation, or university, and "Ford" for SeedTerm2 can refer to a car company, an automobile pioneer, or a U.S. president. But when the two terms co-occur in a text that matches the pattern *"Presidents such as Ford and <X>"*, the text will almost certainly refer to country presidents.

The first stage involves a series of repeated replacements of SeedTerm2 by newly-learned terms in order to generate even more seed terms. That is, each new basic-level concept or instance is rotated into the pattern (becoming a new SeedTerm2) in a bootstrapping cycle that Kozareva et al. called *reckless bootstrapping*. This procedure is implemented as exhaustive breadth-first search, and iterates until no new terms are harvested. The harvested terms are incorporated in a *Hyponym Pattern Linkage Graph (HPLG)* $G = (V, E)$, where each vertex $v \in V$ is a candidate term and each edge $(u, v) \in E$ indicates that term v was generated by term u. A term u is ranked by *Out-Degree(u)* $= \frac{\sum_{\forall (u,v) \in E} w(u,v)}{|V|-1}$, which represents the weighted sum of u's outgoing edges normalized by the total number of other nodes in the graph. Intuitively, a term ranks highly if it is frequently discovering many different terms during the reckless bootstrapping cycle. This method is very productive, harvesting a constant stream of new terms for basic-level concepts or instances when the taxonomy below the initial root concept SeedTerm1 is extensive (such as for Animals or People).

[1] Strictly speaking, our lowest-level concepts can be instances, basic-level concepts, or concepts below the basic level (e.g., *dachsund*). But for the sake of simplicity we will refer to our lowest-level terms as basic-level concepts and instances.

Stage 2: Intermediate level concept collection: Going beyond (Kozareva et al., 2008), we next apply the Doubly-Anchored Pattern in the 'backward' direction (DAP^{-1}), for any two seed terms representing basic-level concepts or instances:

DAP^{-1}: *<X> such as [SeedTerm1] and [SeedTerm2]*

which harvests a set of concepts, most of them intermediate between the basic level or instance and the initial higher-level seed.

This second stage (DAP^{-1}) has not yet been described in the literature. It proceeds analogously. For pairs of basic-level concepts or instances below the root concept that were found during the first stage, we instantiate DAP^{-1} and issue a new web query. For example, if the term "cats" was harvested by DAP in *"Animals such as dogs and <X>"*, then the pair $< dogs, cats >$ forms the new Web query *"<X> such as dogs and cats"*. We extract up to 2 consecutive nouns from the *<X>* position.

This procedure yields a large number of discovered concepts, but they cannot all be used for further bootstrapping. In addition to practical limitations (such as limits on web querying), many of them are too general–more general than the initial root concept–and could derail the bootstrapping process by introducing terms that stray every further away from the initial root concept. We therefore rank the harvested terms based on the likelihood that they will be productive if they are expanded in the next cycle. Ranking is based on two criteria: (1) the concept should be prolific (i.e., produce many lower-level concepts) in order to keep the bootstrapping process energized, and (2) the concept should be subordinate to the root concept, so that the process stays within the targeted part of the search space.

To perform ranking, we incorporate both the harvested concepts and the basic-level/instance pairs into a *Hypernym Relation Graph (HRG)*, which we define as a bipartite graph $HRG = (V, E)$ with two types of vertices. One set of vertices represents the concepts *(the category vertices (V_c)*, and a second set of vertices represents the basic-level/instance pairs that produced the concepts (*the member pair vertices (V_{mp})*). We create an edge $e(u, v) \in E$ between $u \in V_c$ and $v \in V_{mp}$ when the concept represented by u was harvested by the basic-level/instance pair represented by v, with the weight of the edge defined as the number of times that the lower pair found the concept on the web.

We use the Hypernym Relation Graph to rank the intermediate concepts based on each node's *In-Degree*, which is the sum of the weights on the node's incoming edges. Formally, *In-Degree(u)* = $\sum_{\forall(u,v)\in E} w(u, v)$. Intuitively, a concept will be ranked highly if it was harvested by many different combinations of basic-level/instance terms.

However, this scoring function does not determine whether a concept is more or less general than the initial root concept. For example, when harvesting animal categories, the system may learn the word *"species"*, which is a very common term associated with animals, but also applies to non-animals such as plants. To prevent the inclusion of overgeneral terms and constrain the search to remain 'below' the root concept, we apply a *Concept Positioning Test (CPT)*: We issue the following two web queries:

(a) *Concept such as RootConcept and <X>*
(b) *RootConcept such as Concept and <X>*

If (b) returns more web hits than (a), then the concept passes the test, otherwise it fails. The first (most highly ranked) concept that passes CPT becomes the new seed concept for the next bootstrapping cycle. In principle, we could use all the concepts that pass the CPT for bootstrapping[2]. However, for practical reasons (primarily limitations on web querying), we run the algorithm for 10 iterations.

3 Related Work

Many algorithms have been developed to automatically acquire semantic class members using a variety of techniques, including co-occurrence statistics (Riloff and Shepherd, 1997; Roark and Charniak, 1998), syntactic dependencies (Pantel and Ravichandran, 2004), and lexico-syntactic patterns (Riloff and Jones, 1999; Fleischman and Hovy, 2002; Thelen and Riloff, 2002).

The work most closely related to ours is that of (Hearst, 1992) who introduced the idea of applying *hyponym patterns* to text, which explicitly identify a hyponym relation between two terms (e.g.,

[2]The number of ranked concepts that pass CPT changes in each iteration. Also, the wildcard * is important for counts, as can be verified with a quick experiment using Google.

"such authors as <X>"). In recent years, several researchers have followed up on this idea using the web as a corpus. (Pasca, 2004) applies lexico-syntactic hyponym patterns to the Web and use the contexts around them for learning. KnowItAll (Etzioni et al., 2005) applies the hyponym patterns to extract instances from the Web and ranks them by relevance using mutual information. (Kozareva et al., 2008) introduced a bootstrapping scheme using the doubly-anchored pattern (DAP) that is guided through graph ranking. This approach reported a significant improvement from 5% to 18% over approaches using singly-anchored patterns like those of (Pasca, 2004) and (Etzioni et al., 2005).

(Snow et al., 2005) describe a dependency path based approach that generates a large number of weak hypernym patterns using pairs of noun phrases present in WordNet. They build a classifier using the different hypernym patterns and find among the highest precision patterns those of (Hearst, 1992). Snow et al. report performance of 85% precision at 10% recall and 25% precision at 30% recall for 5300 hand-tagged noun phrase pairs. (McNamee et al., 2008) use the technique of (Snow et al., 2005) to harvest the hypernyms of the proper names. The average precision on 75 automatically detected categories is 53%. The discovered hypernyms were intergrated in a Question Answering system which showed an improvement of 9% when evaluated on a TREC Question Answering data set.

Recently, (Ritter et al., 2009) reported hypernym learning using (Hearst, 1992) patterns and manually tagged common and proper nouns. All hypernym candidates matching the pattern are acquired, and the candidate terms are ranked by mutual information. However, they evaluate the performance of their hypernym algorithm by considering only the top 5 hypernyms given a basic-level concept or instance. They report 100% precision at 18% recall, and 66% precision at 72% recall, considering only the top-5 list. Necessarily, using all the results returned will result in lower precision scores. In contrast to their approach, our aim is to first acquire automatically with minimal supervision the basic-level concepts for given root concept. Thus, we almost entirely eliminate the need for humans to provide hyponym seeds. Second, we evaluate the performance of our approach not by measuring the top-ranked 5 hypernyms given a basic-level concept, but considering all harvested hypernyms of the concept.

Unlike (Etzioni et al., 2005), (Pasca, 2007) and (Snow et al., 2005), we learn both instances and concepts simultaneously.

Some researchers have also worked on reorganizing, augmenting, or extending semantic concepts that already exist in manually built resources such as WordNet (Widdows and Dorow, 2002; Snow et al., 2005) or Wikipedia (Ponzetto and Strube, 2007). Work in automated ontology construction has created lexical hierarchies (Caraballo, 1999; Cimiano and Volker, 2005; Mann, 2002), and learned semantic relations such as meronymy (Berland and Charniak, 1999; Girju et al., 2003).

4 Evaluation

The root concepts discussed in this paper are Animals and People, because they head large taxonomic structures that are well-represented in WordNet. Throughout these experiments, we used as the initial SeedTerm2 *lions* for Animals and *Madonna* for People (by specifically choosing a proper name for People we force harvesting down to the level of individual instances). To collect data, we submitted the DAP patterns as web queries to Google, retrieved the top 1000 web snippets per query, and kept only the unique ones. In total, we collected 1.1 GB of snippets for Animals and 1.5 GB for People. The algorithm was allowed to run for 10 iterations.

The algorithm learns a staggering variety of terms that is much more diverse than we had anticipated. In addition to many basic-level concepts or instances, such as *dog* and *Madonna* respectively, and many intermediate concepts, such as *mammals, pets*, and *predators*, it also harvested categories that clearly seemed useful, such as *laboratory animals, forest dwellers*, and *endangered species*. Many other harvested terms were more difficult to judge, including *bait, allergens, seafood, vectors, protein*, and *pests*. While these terms have an obvious relationship to Animals, we have to determine whether they are legitimate and valuable subconcepts of Animals.

A second issue involves relative terms that are hard to define in an absolute sense, such as *native animals* and *large mammals*.

A complete evaluation should answer the following three questions:

- Precision: What is the correctness of the harvested concepts? (How many of them are simply wrong, given the root concept?)
- Recall: What is the coverage of the harvested concepts? (How many are missing, below a given root concept?)
- How correct is the taxonomic structure learned?

Given the number and variety of terms obtained, we initially decided that an automatic evaluation against existing resources (such as WordNet or something similar) would be inadequate because they do not contain many of our harvested terms, even though many of these terms are clearly sensible and potentially valuable. Indeed, the whole point of our work is to learn concepts and taxonomies that go above and beyond what is currently available.

However, it is necessary to compare with *some*thing, and it is important not to skirt the issue by conducting evaluations that measure subsets of results, or that perhaps may mislead. We therefore decided to compare our results against WordNet *and* to have human annotators judge as many results as we could afford (to obtain a measure of Precision and the legitimate extensions beyond WordNet).

Unfortunately, it proved impossible to measure Recall against WordNet, because this requires ascertaining the number of synsets in WordNet between the root and its basic-level categories. This requires human judgment, which we could not afford. We plan to address this question in future work. Also, assessing the correctness of the learned taxonomy structure requires the manual assessment of each classification link proposed by the system that is not already in WordNet, a task also beyond our budget to complete in full. Some results—for just basic-level terms and intermediate concepts, but not among intermediate-level concepts–are shown in Section 4.3.

We provide Precision scores using the following measures, where *terms* refers to the harvested terms:

$$Pr_{WN} = \frac{\#terms\ found\ in\ WordNet}{\#terms\ harvested\ by\ system}$$

$$Pr_{H} = \frac{\#terms\ judged\ correct\ by\ human}{\#terms\ harvested\ by\ system}$$

$$NotInWN = \#terms\ judged\ correct\ by\ human\ but\ not\ in\ WordNet$$

We conducted three sets of experiments. **Experiment 1** evaluates the results of using DAP to learn basic-level concepts for Animals and instances for People. **Experiment 2** evaluates the results of using DAP^{-1} to harvest intermediate concepts between each root concept and its basic-level concepts or instances. **Experiment 3** evaluates the taxonomy structure that is produced via the links between the instances and intermediate concepts.

4.1 Experiment 1: Basic-Level Concepts and Instances

In this section we discuss the results of harvesting the basic-level Animal concepts and People instances. The bootstrapping algorithm ranks the harvested terms by their *Out-Degree* score and considers as correct only those with *Out-Degree* > 0. In ten iterations, the bootstrapping algorithm produced 913 Animal basic-level concepts and 1,344 People instances that passed this *Out-Degree* criterion.

4.1.1 Human Evaluation

The harvested terms were labeled by human judges as either correct or incorrect with respect to the root concept. Table 1 shows the Precision of the top-ranked N terms, with N shown in increments of 100. Overall, the Animal terms yielded 71% (649/913) Precision and the People terms yielded 95% Precision (1,271/1,344). Figure 1 shows that higher-ranked Animal terms are more accurate than lower-ranked terms, which indicates that the scoring function did its job. For People terms, accuracy was very high throughout the ranked list. Overall, these results show that the bootstrapping algorithm generates a large number of correct instances of high quality.

4.1.2 WordNet Evaluation

Table 1 shows a comparison of the harvested terms against the terms present in WordNet. Note that the Precision measured against WordNet (Pr_{WN}) for People is dramatically different from the Precision based on human judgments (Pr_H). This can be explained by looking at the *NotInWN* column, which shows that 48 correct Animal terms

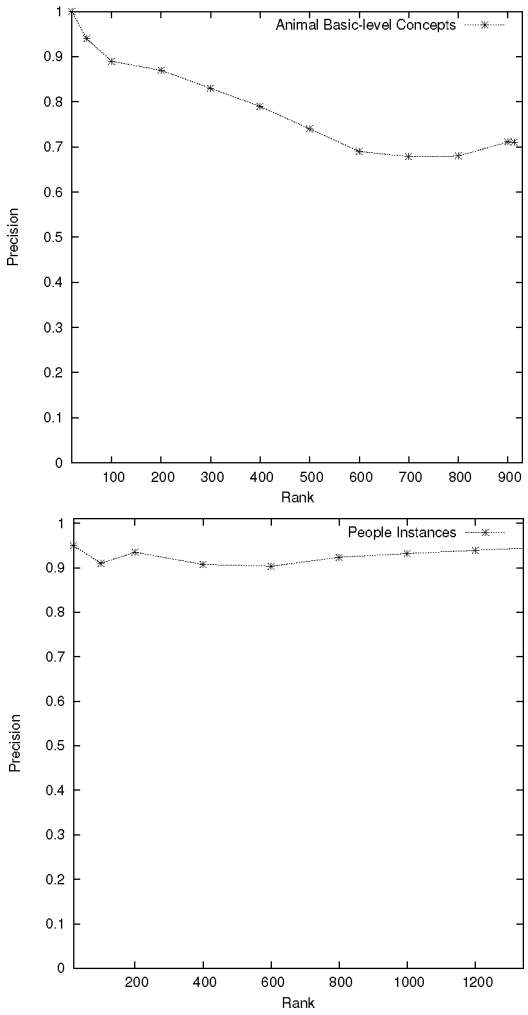

Figure 1: *Ranked Basic-Concepts and Instances.*

and 986 correct People instances are not present in WordNet (primarily, for people, because WordNet contains relatively few proper names). These results show that there is substantial room for improvement in WordNet's coverage of these categories. For Animals, the precision measured against WordNet is actually higher than the precision measured by human judges, which may indicate that the judges failed to recognize some correct animal terms.

	Pr_{WN}	Pr_{H}	$NotInWN$
Animal	.79	.71	48
People	.23	.95	986

Table 1: *Instance Evaluation.*

4.1.3 Evaluation against Prior Work

To assess how well our algorithm compares with previous semantic class learning methods, we compared our results to those of (Kozareva et al., 2008). Our work was inspired by that approach–in fact, we use that previous algorithm as the first step of our bootstrapping process. The novelty of our approach is the insertion of an additional bootstrapping stage that iteratively learns new intermediate concepts using DAP^{-1} and the Concept Positioning Test, followed by the subsequent use of the newly learned intermediate concepts in DAP to expand the search space beyond the original root concept. This leads to the discovery of additional basic-level terms or instances, which are then recycled in turn to discover new intermediate concepts, and so on.

Consequently, we can compare the results produced by the first iteration of our algorithm (before intermediate concepts are learned) to those of (Kozareva et al., 2008) for the Animal and People categories, and then compare again after 10 bootstrapping iterations of intermediate concept learning. Figure 2 shows the number of harvested concepts for Animals and People after each bootstrapping iteration. Bootstrapping with intermediate concepts produces nearly 5 times as many basic-level concepts and instances than (Kozareva et al., 2008) obtain, while maintaining similar levels of precision.

The intermediate concepts help so much because they steer the learning process into new (yet still correct) regions of the search space after each iteration. For instance, in the first iteration, the pattern *"animals such as lions and *"* harvests about 350 basic-level concepts, but only animals that are mentioned in conjunction with lions are learned. Of these, animals typically quite different from lions, such as grass-eating kudu, are often not discovered.

However, in the second iteration, the intermediate concept Herbivore is chosen for expansion. The pattern *"herbivore such as antelope and *"* discovers many additional animals, including *kudu*, that co-occur with *antelope* but do not co-occur with *lions*.

Table 2 shows examples of the 10 top-ranked basic-level concepts and instances that were learned for 3 randomly-selected intermediate Animal and People concepts (*IConcepts*) that were acquired during bootstrapping. In the next section, we present an

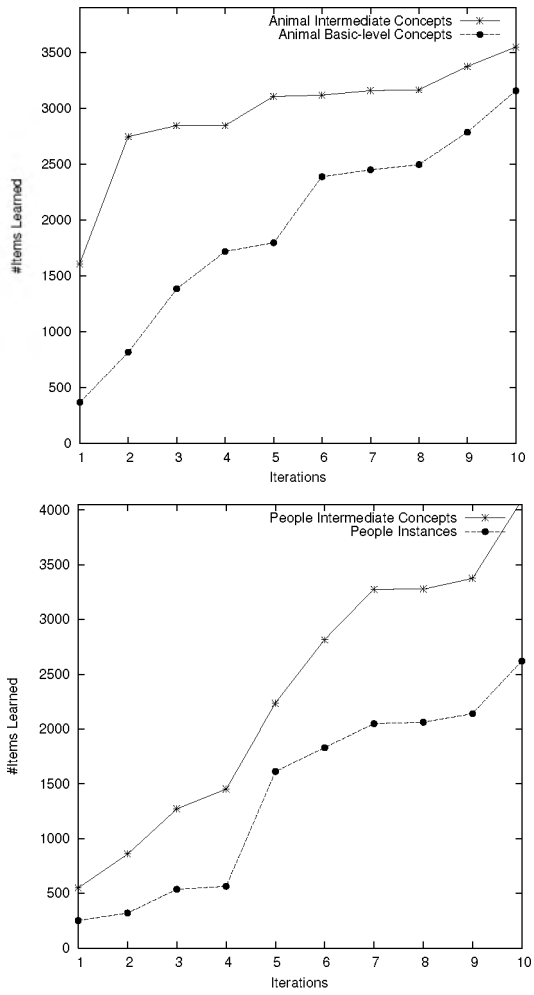

Figure 2: *Learning Curves.*

evaluation of the intermediate concept terms.

4.2 Experiment 2: Intermediate Concepts

In this section we discuss the results of harvesting the intermediate-level concepts. Given the variety of the harvested results, manual judgment of correctness required an in-depth human annotation study. We also compare our harvested results against the concept terms in WordNet.

4.2.1 Human Evaluation

We hired 4 annotators (undergraduates at a different institution) to judge the correctness of the intermediate concepts. We created detailed annotation guidelines that define 14 annotation labels for each of the Animal and People classes, as shown in Table 3. The labels are clustered into 4 major types: *Correct, Borderline, BasicConcept, and NotConcept*. The details of our annotation guidelines, the reasons for the intermediate labels, and the annotation study can be found in (Kozareva et al., 2009).

PEOPLE	
IConcept	*Instances*
Dictators:	Adolf Hitler, Joseph Stalin, Benito Mussolini, Lenin, Fidel Castro, Idi Amin, Slobodan Milosevic, Hugo Chavez, Mao Zedong, Saddam Hussein
Celebrities:	Madonna, Paris Hilton, Angelina Jolie, Britney , Spears, Tom Cruise, Cameron Diaz, Bono, Oprah Winfrey, Jennifer Aniston, Kate Moss
Writers:	William Shakespeare, James Joyce, Charles Dickens, Leo Tolstoy, Goethe, Ralph Waldo Emerson, Daniel Defoe, Jane Austen, Ernest Hemingway, Franz Kafka

ANIMAL	
IConcept	*Basic-level Terms*
Crustacean:	shrimp, crabs, prawns, lobsters, crayfish, mysids, decapods, marron, ostracods, yabbies
Primates:	baboons, monkeys, chimpanzees, apes, marmosets, chimps, orangutans, gibbons, tamarins, bonobos
Mammal:	mice, whales, seals, dolphins, rats, deer, rabbits, dogs, elephants, squirrels

Table 2: *Learned People and Animals Terms.*

ANIMAL

TYPE	LABEL	EXAMPLES
Correct	GeneticAnimal	*reptile,mammal*
	BehavioralByFeeding	*predator, grazer*
	BehaviorByHabitat	*saltwater mammal*
	BehaviorSocialIndiv	*herding animal*
	BehaviorSocialGroup	*herd, pack*
	MorphologicalType	*cloven-hoofed animal*
	RoleOrFunction	*pet, parasite*
Borderline	NonRealAnimal	*dragons*
	EvaluativeTerm	*varmint, fox*
	OtherAnimal	*critter, fossil*
BasicConcept	BasicAnimal	*dog, hummingbird*
NotConcept	GeneralTerm	*model, catalyst*
	NotAnimal	*topic, favorite*
	GarbageTerm	*brates, mals*

PEOPLE

TYPE	LABEL	EXAMPLES
Correct	GeneticPerson	*Caucasian, Saxon*
	NonTransientEventRole	*stutterer, gourmand*
	TransientEventRole	*passenger, visitor*
	PersonState	*dwarf, schizophrenic*
	FamilyRelation	*aunt, mother*
	SocialRole	*fugitive, hero*
	NationOrTribe	*Bulgarian, Zulu*
	ReligiousAffiliation	*Catholic, atheist*
Borderline	NonRealPerson	*biblical figures*
	OtherPerson	*colleagues, couples*
BasicConcept	BasicPerson	*child, woman*
	RealPerson	*Barack Obama*
NotConcept	GeneralTerm	*image, figure*
	NotPerson	*books, events*

Table 3: *Intermediate Concept Annotation Labels*

We measured pairwise inter-annotator agreement across the four labels using the Fleiss kappa (Fleiss, 1971). The κ scores ranged from 0.61–0.71 for Animals (average κ=0.66) and from 0.51–0.70 for People (average κ=0.60). These agreement scores seemed good enough to warrant using these human judgments to estimate the accuracy of the algorithm.

The bootstrapping algorithm harvested 3,549 Animal and 4,094 People intermediate concepts in ten iterations. After *In-Degree* ranking was applied,

we chose a random sample of intermediate concepts with frequency over 1, which was given to four human judges for annotation. Table 4 summarizes the labels assigned by the four annotators (A_1 – A_4). The top portion of Table 4 shows the results for all the intermediate concepts (437 Animal terms and 296 People terms), and the bottom portion shows the results only for the concepts that passed the Concept Positioning Test (187 Animal terms and 139 People terms). Accuracy is computed in two ways: **Acc1** is the percent of intermediate concepts labeled as *Correct*; **Acc2** is the percent of intermediate concepts labeled as either *Correct* or *Borderline*.

Without the CPT, accuracies range from 53–66% for Animals and 75–85% for People. After applying the CPT, the accuracies increase to 71–84% for animals and 82–94% for people. These results confirm that the Concept Positioning Test is effective at removing many of the undesirable terms. Overall, these results demonstrate that our algorithm produced many high-quality intermediate concepts, with good precision.

Figure 3 shows accuracy curves based on the rankings of the intermediate concepts (based on In-Degree scores). The CPT clearly improves accuracy even among the most highly ranked concepts. For example, the **Acc1** curves for animals show that nearly 90% of the top 100 intermediate concepts were correct after applying the CPT, whereas only 70% of the top 100 intermediate concepts were correct before. However, the CPT also eliminates many desirable terms. For People, the accuracies are still relatively high even without the CPT, and a much larger set of intermediate concepts is learned.

	Animals				People			
	A_1	A_2	A_3	A_4	A_1	A_2	A_3	A_4
Correct	246	243	251	230	239	231	225	221
Borderline	42	26	22	29	12	10	6	4
BasicConcept	2	8	9	2	6	2	9	10
NotConcept	147	160	155	176	39	53	56	61
Acc1	.56	.56	.57	.53	.81	.78	.76	.75
Acc2	.66	.62	.62	.59	.85	.81	.78	.76
	Animals after CPT				**People after CPT**			
	A_1	A_2	A_3	A_4	A_1	A_2	A_3	A_4
Correct	146	133	144	141	126	126	114	116
Borderline	11	15	9	13	6	2	2	0
BasicConcept	2	8	9	2	0	1	7	7
NotConcept	28	31	25	31	7	10	16	16
Acc1	.78	.71	.77	.75	.91	.91	.82	.83
Acc2	.84	.79	.82	.82	.95	.92	.83	.83

Table 4: *Human Intermediate Concept Evaluation.*

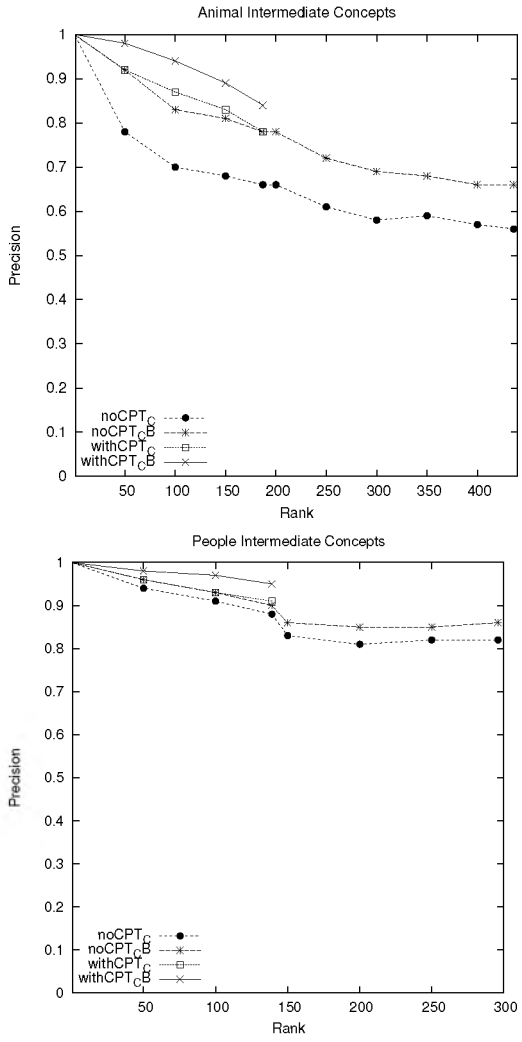

Figure 3: *Intermediate Concept Precision at Rank N.*

4.2.2 WordNet Evaluation

We also compared the intermediate concepts harvested by the algorithm to the contents of WordNet. The results are shown in Table 5. WordNet contains 20% of the Animal concepts and 51% of the People concepts learned by our algorithm, which confirms that many of these concepts were considered to be valuable taxonomic terms by the WordNet developers. However, our human annotators judged 57% of the Animal and 84% of the People concepts to be correct, which suggests that our algorithm generates a substantial number of additional concepts that could be used to enrich taxonomic structure in WordNet.

	Pr_{WN}	Pr_H	NotInWN
Animal	.20 (88/437)	.57 (248/437)	204
People	.51 (152/296)	.85 (251/296)	108

Table 5: *WordNet Intermediate Concept Evaluation.*

4.3 Experiment 3: Taxonomic Links

In this section we evaluate the classification (taxonomy) that is learned by evaluating the links between the intermediate concepts and the basic-level concept/instance terms. That is, when our algorithm claims that *isa(X,Y)*, how often is *X* truly a subconcept of *Y*? For example, *isa(goat, herbivore)* would be correct, but *isa(goat, bird)* would not. Again, since WordNet does not contain all the harvested concepts, we conduct both a manual evaluation and a comparison against WordNet.

4.3.1 Manual and WordNet Evaluations

Creating and evaluating the full taxonomic structure between the root and the basic-level or instance terms is future work. Here we evaluate simply the accuracy of the taxonomic links between basic-level concepts/instances and intermediate concepts as harvested, but not between intermediate concepts. For each pair, we extracted all harvested links and determined whether the same links appear in WordNet. The links were also given to human judges. Table 6 shows the results.

ISA	Pr_{WN}	Pr_H	NotInWN
Animal	.47(912/1940)	.88 (1716/1940)	804
People	.23 (318/908)	.94 (857/908)	539

Table 6: *WordNet Taxonomic Evaluation.*

The results show that WordNet lacks nearly half of the taxonomic relations that were generated by the algorithm: 804 Animal and 539 People links.

5 Conclusion

We describe a novel extension to the DAP approach for discovering basic-level concepts or instances and their superconcepts given an initial root concept. By appropriate filling of different positions in DAP, the algorithm alternates between 'downward' and 'upward' learning. A key resulting benefit is that each new intermediate-level term acquired restarts harvesting in a new region of the concept space, which allows previously unseen concepts to be discovered with each bootstrapping cycle.

We also introduce the *Concept Positioning Test*, which serves to confirm that a harvested concept falls into the desired part of the search space relative to either a superordinate or subordinate concept in the growing taxonomy, before it is selected for further harvesting using the DAP.

These algorithms can augment other term harvesting algorithms recently reported. But in order to compare different algorithms, it is important to compare results to a standard. WordNet is our best candidate at present. But WordNet is incomplete. Our results include a significantly large number of instances of People (which WordNet does not claim to cover), a number comparable to the results of (Etzioni et al., 2005; Pasca, 2007; Ritter et al., 2009). Rather surprisingly, our results also include a large number of basic-level and intermediate concepts for Animals that are not present in WordNet, a category WordNet is actually fairly complete about. These numbers show clearly that it is important to conduct manual evaluation of term harvesting algorithms in addition to comparing to a standard resource.

Acknowledgments

This research was supported in part by grants from the National Science Foundation (NSF grant no. IIS-0429360), and the Department of Homeland Security, ONR Grant numbers N0014-07-1-0152 and N00014-07-1-0149. We are grateful to the annotators at the University of Pittsburgh who helped us evaluate this work: Jay Fischer, David Halpern, Amir Hussain, and Taichi Nakatani.

References

M. Banko, M. Cafarella, S. Soderland, M. Broadhead, and O.Etzioni. 2007. Open information extraction from the web. In *Proceedings of International Joint Conference on Artificial Itelligence*, pages 2670–2676.

M. Berland and E. Charniak. 1999. Finding Parts in Very Large Corpora. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics.*

S. Caraballo. 1999. Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 120–126.

P. Cimiano and J. Volker. 2005. Towards large-scale, open-domain and ontology-based named entity classification. In *Proceeding of RANLP-05*, pages 166–172.

O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. 2005. Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence*, 165(1):91–134, June.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. May.

M.B. Fleischman and E.H. Hovy. 2002. Fine grained classification of named entities. In *Proceedings of the COLING conference*, August.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

R. Girju, A. Badulescu, and D. Moldovan. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In *HLT-NAACL*.

M. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING*, pages 539–545.

Z. Kozareva, E. Riloff, and E. Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. In *Proceedings of ACL-08: HLT*, pages 1048–1056. Association for Computational Linguistics.

Z. Kozareva, E. Hovy, and E. Riloff. 2009. Learning and evaluating the content and structure of a term taxonomy. In *AAAI-09 Spring Symposium on Learning by Reading and Learning to Read*.

G. Mann. 2002. Fine-grained proper noun ontologies for question answering. In *COLING-02 on SEMANET*, pages 1–7.

P. McNamee, R. Snow, P. Schone, and J. Mayfield. 2008. Learning named entity hyponyms for question answering. In *Proceedings of the Third International Joint Conference on Natural Language Processing*.

P. Pantel and D. Ravichandran. 2004. Automatically labeling semantic classes. In *HLT-NAACL*, pages 321–328.

M. Pasca. 2004. Acquisition of categorized named entities for web search. In *Proceedings of CIKM*, pages 137–145.

M. Pasca. 2007. Weakly-supervised discovery of named entities using web search queries. In *CIKM*, pages 683–690.

S. Ponzetto and M. Strube. 2007. Deriving a large scale taxonomy from wikipedia. In *Proceedings of the 22nd National COnference on Artificial Intelligence (AAAI-07)*, pages 1440–1447.

E. Riloff and R. Jones. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*.

E. Riloff and J. Shepherd. 1997. A Corpus-Based Approach for Building Semantic Lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124.

A. Ritter, S. Soderland, and O. Etzioni. 2009. What is this, anyway: Automatic hypernym discovery. In *Proceedings of AAAI-09 Spring Symposium on Learning by Reading and Learning to Read*, pages 88–93.

B. Roark and E. Charniak. 1998. Noun-phrase Co-occurrence Statistics for Semi-automatic Semantic Lexicon Construction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 1110–1116.

E. Rosch, 1978. *Principles of Categorization*, pages 27–48.

R. Snow, D. Jurafsky, and A. Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *NIPS*.

M. Thelen and E. Riloff. 2002. A Bootstrapping Method for Learning Semantic Lexicons Using Extraction Pattern Contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 214–221.

D. Widdows and B. Dorow. 2002. A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7.

J. Yi and W. Niblack. 2005. Sentiment mining in webfountain. In *ICDE '05: Proceedings of the 21st International Conference on Data Engineering*, pages 1073–1083.