Lang Resources & Evaluation (2013) 47:859–890
DOI 10.1007/s10579-013-9229-0

ORIGINAL PAPER

# Tailoring the automated construction of large-scale taxonomies using the web

**Zornitsa Kozareva · Eduard Hovy**

Published online: 7 May 2013

**Abstract** It has long been a dream to have available a single, centralized, semantic thesaurus or terminology taxonomy to support research in a variety of fields. Much human and computational effort has gone into constructing such resources, including the original WordNet and subsequent wordnets in various languages. To produce such resources one has to overcome well-known problems in achieving both wide coverage and internal consistency within a single wordnet and across many wordnets. In particular, one has to ensure that alternative valid taxonomizations covering the same basic terms are recognized and treated appropriately. In this paper we describe a pipeline of new, powerful, minimally supervised, automated algorithms that can be used to construct terminology taxonomies and wordnets, in various languages, by harvesting large amounts of online domain-specific or general text. We illustrate the effectiveness of the algorithms both to build localized, domain-specific wordnets and to highlight and investigate certain deeper ontological problems such as parallel generalization hierarchies. We show shortcomings and gaps in the manually-constructed English WordNet in various domains.

**Keywords** Hyponym and hypernym learning · Text mining · Ontology induction · Wordnet evaluation

## 1 Introduction

Even before the appearance of the original WordNet (Miller 1995; Fellbaum 1998), but especially since then, there has been a great deal of effort in (semi-

Z. Kozareva (✉) · E. Hovy
USC Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292-6695, USA
e-mail: kozareva@isi.edu

E. Hovy
e-mail: hovy@isi.edu

)automatically creating terminology taxonomies and wordnets of English and other languages. The hope is that such resources will provide ontological and lexicographical information useful for automated text processing. Several types of alternatives have appeared, including large and elaborated Artificial Intelligence ontologies such as CYC (Lenat and Guha 1990), terminology taxonomies such as WordNet covering English in general (Fellbaum 1998) or other languages (Pease et al. 2008; Vossen et al. 2008; Atserias et al. 2004; Robkop et al. 2010; Navigli and Ponzetto 2012), large domain-oriented terminology banks covering medical and other terminology (UMLS www.nlm.nih.gov/research/umls/, Velardi et al. 2008), automatically harvested instance repositories such as YAGO (Suchanek et al. 2007) and NELL (Mitchell et al. 2009), and numerous smaller domain-specialized terminology taxonomies and ontologies.

However, despite considerable effort, no satisfactory wordnet exists today, even for English. The reasons include:

- the difficulty to obtain adequate coverage over all words of the language/domain,
- the complexity of ontological decisions about word senses and sense groupings,
- the difficulty to build consistent subsumption/generalization hierarchies using hypernym/hyponym relations,
- the difficulty to obtain additional inter-term relations.

These problems are deep and not trivially resolved via algorithms alone. For example, the ontological decision—which superconcept/hypernym to select for a given term?—may have several correct parallel answers, as we describe in Sect. 5.2 Specialized algorithms that go beyond traditional superclass categorization via patterns or glossary definitions may need to be developed. In general, these problems require careful thought, and most of them (at this point) also require considerable human effort, to collect, sort, and link terms.

All ontology and wordnet construction efforts run into the problem of internal consistency once a certain degree of coverage and internal richness is achieved. It seems impossible to create a single semantic knowledge repository/thesaurus that is simultaneously rich in detail, very large (say, over 100,000 different concepts), and internally consistent. While to an initial repository one can relatively easily add more instances of existing concepts (say, more individual humans under the concept *Singer*), it seems very difficult to continue to add additional concepts and organize them all relative to one another in ways that support uniform inference across the whole repository. Usually, concepts are organized into taxonomies of increasing specificity; the Animal Kingdom provides a good example. But there are many conceptualizations of animals that do not fit neatly into a single taxonomy. The kind of organization that would support, for example, the types *Dog, Cat, Mammal, Pet, Carnivore, Domesticated Animal, Endoskeleton, Animal*, and so on, is probably a set of parallel and interlinked taxonomies. But this strategy does not really work for the set of Emotion concepts, or for most of the Events.

The problem is exacerbated when one attempts to develop a single semantic model that supports multiple languages. Even closely related languages such as Dutch and English or Spanish and Italian exhibit relative incompatibilities—not just lexical (and perhaps conceptual gaps), but actually different partitioning of the same semantic field into apparently different conceptualizations.

What can be done about this?

Ideally, one would solve the conceptual problems and then develop automated methods to (help) construct the desired results. But sometimes the conceptual problems are most apparent only when one has available a large number of terms to work with. Therefore, automated algorithms that perform some of these tasks, such as collecting many terms and organizing them, and that can be interleaved with human analysis and correction, are highly desirable.

Unfortunately, to date, automated ontology construction work has not fully resolved these problems. This may be due to the ambitious nature of previous attempts to try to solve too many of the problems all at once (see for example Snow et al. 2006). Rather, we believe it is more effective to break the problem into a series of smaller steps, and to develop algorithms for each step, and also to try to localize some of the harder conceptual/ontological problems within individual steps rather than across the whole process.

The most straightforward step-wise procedure is to first collect the terms that will constitute the wordnet, then to create a single central backbone structure, e.g., a generalization taxonomy or DAG, of core conceptualizations using hypernym/hyponym relations, and then to interlink the terms using other relations. These steps can be performed manually, automatically, or in mixed mode. Addressing multiple languages, one can try to create a single multilingual wordnet, a set of parallel and interlinked wordnets, or simply a set of independent unlinked wordnets. An early attempt to create the first option, using a hybridized multilingual Upper Model (Bateman et al. 1989) to help fuse the terms from various languages, (Hovy and Nirenburg 1992) failed. A much more substantive attempt to create the second was the EuroWordNet project (Vossen et al. 1998), in which the cross-linking was achieved using the so-called Inter-Linking Index ILI. Even though EuroWordNet focused on just a handful of relatively closely related languages (English Dutch, Italian, Spanish, and later German), the ILI approach still posed problems. As a result, this approach is not used today to interlink the various language-based WordNets being built around the world in the Global WordNet endeavor (Pease et al. 2008; Vossen et al. 2008).

The third alternative is to first create independent domain-specific wordnets in one or more languages and then fuse them to the degree possible. For this option, algorithms that can rapidly, with minimal supervision, create a new localized terminology taxonomy around one or more starting terms, given any new corpus in any language, would be most helpful. One can then attempt to fuse them with due consideration to and exploitation of the differences encountered across neighboring wordnets and/or across languages.

In this paper, we describe a series of simple term harvesting, taxonomization, and interlinking algorithms that require very little supervision yet deliver high precision and wide coverage, given online texts in any language. The rest of the paper is organized as follows. Section 2 outlines basic terminology and the general approach. Section 3 reviews related work. Section 4 describes the employed lexico-syntactic pattern. Section 5 describes the core methods for knowledge extraction, which are followed in Sect. 6 by the taxonomization algorithm. Section 7 provides a detailed human based evaluation of the harvested hyponym, hypernym terms and is-a relations for four different domains of interests. We conduct a comparative study against WordNet and existing knowledge harvesting methods, and discuss the results in Sect. 8 Finally, we conclude in Sect. 9.

## 2 Terminology

### 2.1 Basic terminology

Prior to introducing our work, we define some basic terminology that is used in the paper, since these terms can be interpreted differently by different scientific communities.

- **term**: A single English word (or possibly a two-word fixed phrase, such as "opera singer") that denotes a single concept.
- **seed term**: A term that is employed at the outset of the harvesting process. Usually, the seed term is selected by a human.
- **concept**: An item in the classification taxonomy we are building.[1]
- **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. Example: *animal, plant, people*.
- **low-level concept**: A concept at a fairly low level in the taxonomy, to which many others are eventually learned to be supertype. Typically the concept can be visualized (i.e., one can visualize a dog, but not a mammal) (Rosch 1978). Example: *dog, mountain, Madonna*.
- **intermediate-level concept**: A concept located between the root and the low-level concept. Example: *mammal, shrub, teacher*.
- **classification link**: A link that expresses the subsumption (*is-a*) relation between two concepts. The word from more-specific 'upward' to more general term is called *hypernym* and the opposite, *hyponym*.

### 2.2 Problem formulation

Breaking down the problem of (semi-)automatically creating wordnets into a series of steps, we define our task as knowledge harvesting and knowledge organization procedures.

Figure 1 shows an illustrative example of our task. The algorithm is instantiated with the root concept *animal* and the low-level concept *lion*. The algorithm learns new low-level terms like *tiger, puma, deer, donkey* of class *animal* and then uses these terms to acquire hypernyms like *lion* is-a *vertebrate*, *chordate*, *feline* and *mammal*. To keep the harvesting process within the domain, all harvested terms are validated for subordination with respect to the original root concept *animal*.

[1] For the sake of simplicity in this paper, we will use *term* and *concept* interchangeably.

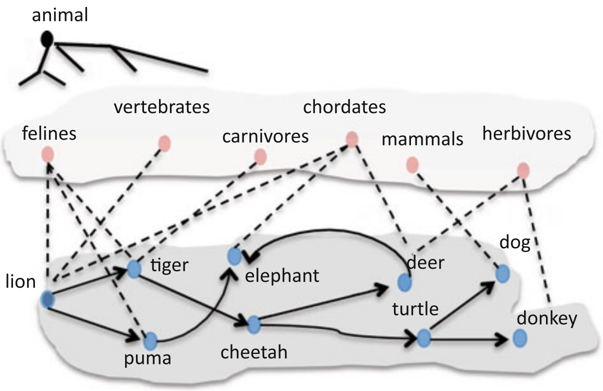

**Fig. 1** Taxonomy induction from scratch

Concepts that do not pass the subordination test are removed, while the rest of the terms are used for taxonomy induction. For instance, *animal* → *chordate* → *vertebrate* → *mammal* → *feline* → *lion*.

**Task Definition** Given a root concept, a low-level concept and a lexico-syntactic pattern: (1) harvest in bootstrapping fashion hyponym and hypernym terms; rerank the terms and filter out all concepts that do not subordinate to the root concept; (2) organize the terms into one or more taxonomies.

## 3 Related work

The first stage of automatic taxonomy induction, term and relation extraction, is relatively well-understood. Early methods (Hearst 1992) have matured to the point of achieving high accuracy (Girju et al. 2003; Pantel and Pennacchiotti 2006; Kozareva et al. 2008). The produced output typically contains flat lists of terms and/or ground instance facts (*lion* is-a *mammal*) and general relation types (*mammal* is-a *animal*).

Most approaches use either clustering or patterns to mine knowledge from structured and unstructured text. Clustering approaches (Lin 1998; Lin and Pantel 2002; Davidov and Rappoport 2006) are fully unsupervised and discover relations that are not directly expressed in text. Their main drawback is that they may or may not produce the term types and granularities useful to the user. In contrast, pattern-based approaches harvest information with high accuracy, but they require a set of seeds and surface patterns to initiate the learning process. These methods are successfully used to collect semantic lexicons (Riloff and Shepherd 1997; Etzioni et al. 2005; Pasca 2004; Kozareva et al. 2008), encyclopedic knowledge (Suchanek et al. 2007; Ponzetto and Navigli 2010; Cuadros and Rigau 2008; Agirre et al. 2004), concept lists (Katz and Lin 2003), and relations between terms, such as hypernyms (Ritter et al. 2009; Hovy et al. 2009) and part-of (Girju et al. 2003; Pantel and Pennacchiotti 2006).

However, simple term lists are not enough to solve many problems involving natural language. Terms may be augmented with information that is required for knowledge-intensive tasks such as textual entailment (Glickman et al. 2005; Szpektor et al. 2008) and question answering (Moldovan et al. 1999). To support inference, (Ritter et al. 2010) learn the selectional restrictions of semantic relations, and (Pennacchiotti and Pantel 2006) ontologize the learned terms using WordNet.

Taxonomizing the terms is a very powerful method to leverage added information. Subordinated terms (hyponyms) inherit information from their superordinates (hypernyms), making it unnecessary to learn all relevant information over and over for every term in the language. But despite many attempts, no 'correct' taxonomization has ever been constructed for the terms of, say, English. Typically, people build term taxonomies (and/or richer structures like ontologies) for particular purposes, using specific taxonomization criteria. Different tasks and criteria produce different taxonomies, even when using the same low-level concepts. This is because most low-level concepts admit to multiple perspectives, while each task focuses on one, or at most two, perspectives at a time. For example, a dolphin is a Mammal (and not a Fish) to a biologist, but is a Fish (and hence not a Mammal) to a fisherman or anyone building or visiting an aquarium. More confusingly, a tiger and a puppy are both Mammals and hence belong close together in a typical taxonomy, but a tiger is a WildAnimal (in the perspective of AnimalFunction) and a JungleDweller (in the perspective of Habitat), while a puppy is a Pet (as function) and a HouseAnimal (as habitat), which would place them relatively far from one another. Attempts at producing a single multi-perspective taxonomy fail due to the complexity of interaction among perspectives, and people are notoriously bad at constructing taxonomies adherent to a single perspective when given terms from multiple perspectives. This issue and the major alternative principles for taxonomization are discussed in (Hovy 2002).

It is therefore not surprising that the second stage of automated taxonomy induction is harder to achieve. Early attempts on acquiring taxonomies from machine reading dictionaries include (Amsler 1981; Wilks et al. 1988; Ide and Veronis 1994; Richardson et al. 1998; Rigau et al. 1998). The most common taxonomy learning approaches start with a reasonably complete taxonomy and then insert the newly learned terms into it, one term at a time (Widdows 2003; Pasca 2004; Snow et al. 2006; Yang and Callan 2009; Hovy 1998). Others (Snow et al. 2006) guide the incremental approach by maximizing the conditional probability over a set of relations, while (Yang and Callan 2009) introduce a taxonomy induction framework which combines the power of surface patterns and clustering through combining numerous heterogeneous features. A third approach mines helpful taxonomization information from glossary definitions and pattern-based extraction, using an approach of graph weighting and pruning (Navigli et al. 2004). Later approaches mix several methods, as in (Navigli et al. 2004).

Our basic approach is to address the ontologizing problem directly by recognizing that, even within a single domain, many terms participate in more than one parallel taxonomies simultaneously. Delivering a complete term taxonomization result for a given subdomain requires determining the possible alternative parallel taxonomies that obtain in that subdomain and then augmenting the basic term taxonomization procedure to localize each harvested term into the appropriate parallel option(s). While we have no automated solution for the first step, we describe in Sect. 5.2 the second. This approach differs fundamentally from earlier approaches in recognizing the need for more than one parallel taxonomy, which complicates the whole process but provides, we hope, answers to some of the pressing unresolved problems surrounding task-specific taxonomizations and perspectives.

Our procedure to organize the harvested terms into a taxonomic structure starting fresh (i.e., without using an initial taxonomic structure) bridges the gap between the term extraction algorithms that focus mainly on harvesting but do not taxonomize, and those that accept a new term and seek to enrich an already existing taxonomy. Our aim is to perform both stages: to extract the terms of a given domain and to induce their taxonomic organization without any initial taxonomic structure and information. This task is challenging because it is not trivial to discover both the hierarchically related and the parallel (perspectival) organizations of concepts. Achieving this goal can provide the research community with the ability to produce taxonomies for domains for which currently there are no existing or manually created ontologies.

In the next section we describe the basic harvesting algorithm, and then show how it is adapted and used to perform the sequence of harvesting and taxonomization steps.

## 4 Doubly-anchored patterns

Our work on knowledge acquisition and taxonomization is inspired by Hearst's observations that sentences contain clues as to their meanings and these can be captured using lexico-syntactic patterns (Hearst 1992).

The most common pattern is the so called singly-anchored pattern (SAP) of the form "$\langle$*seed*$\rangle$ *such as* *", which has one example of the seed term (the anchor) and one open position * for the terms to be learned. Most researchers (Pasca 2004; Etzioni et al. 2005) rely on SAP patterns to harvest hyponyms and hypernyms from the Web, but they report that the patterns run out of steam very quickly.

To surmount this obstacle, (Pasca 2004; Pantel and Pennacchiotti 2006) instantiate the knowledge harvesting algorithm with a handful of seed examples, while (Riloff and Jones 1999; Snow et al. 2005; Etzioni et al. 2005) use multiple variations of the initial lexico-syntactic pattern. Although seed selection seems like a trivial step, (Pantel et al. 2009) show that one must ask human experts to achieve high yield. (Banko 2009) reports that human-based seed selection is quite unrealistic when dealing with an unbounded set of relations.

Interestingly, recent work reports a class of patterns that use only one seed example to learn as much information as the previous approaches. (Kozareva et al. 2008; Hovy et al. 2009) introduce the so-called doubly-anchored pattern (DAP) that has two anchor seed positions "$\langle$*semantic class*$\rangle$ *such as* $\langle$*seed*$\rangle$ *and* *", plus one open position for the terms to be learned. DAP is very reliable because it is instantiated with examples at both ends of the space to be filled (the higher-level concept *type* and an instance (low-level) term *seed*), which mutually disambiguate each other. For example, presidents for *semantic class* can refer to the leader of a country, corporation, or university, and Ford for *seed* 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* *, the text will almost certainly refer to country presidents. The power of DAP also lies in its recursive nature which allows for the newly learned terms on the * position to be automatically replaced into the seed position. In this way the recursion eliminates the need for humans to provide seeds and leads to higher term extraction in comparison to the singly anchored patterns (Kozareva et al. 2008).

We are particularly interested in using the DAPs to learn hyponyms and hypernyms for a given domain of interest. Our main motivation is based on the fact that DAP: (1) has shown to learn terms with higher precision compared to the singly-anchored patterns (Kozareva et al. 2008), (2) uses only one seed instance to discover new and previously unknown terms, (3) acquires knowledge with minimal supervision and (4) can be used as a knowledge extraction and concept positioning mechanism.

## 5 Knowledge harvesting using double-anchored patterns

The first stage of our algorithm concerns knowledge acquisition. We propose a minimally supervised bootstrapping algorithm which uses DAPs in two alternating phrases to learn *hyponyms* and *hypernyms* associated with a given domain of interest. The extracted terms are filtered out and reranked using a concept positioning test (CPT). The general framework of the knowledge harvesting algorithm is shown in Table 1. The final output of this phase is a ranked list of terms and is-a relations.

### 5.1 Hyponym harvesting

The hyponym harvesting phrase (i.e. extraction of concepts located at the low-level of the taxonomy) also incorporates a bootstrapping mechanism on its own, which is instantiated with a *semantic class*, one *seed* term from the *semantic class* and a DAP pattern of the form "$\langle$*semantic class*$\rangle$ *such as* $\langle$*seed*$\rangle$ *and* *", where the * is a placeholder for the terms to be learned. In the first iteration, the *semantic class* is the so called *root* concept, which is a term located higher up in the taxonomy. Root concepts are given by the user and they represent terms like *animal, people, plant* among others. The pattern is submitted to Yahoo! as a web query and all unique snippets matching the query are retrieved. The snippets are part-of-speech tagged with TreeTagger (Schmid 1994) and only the nouns and proper names located on the * position are extracted. From these terms, only the newly learned and previously unexplored ones are used as seeds in the subsequent iteration. The bootstrapping process is implemented as an exhaustive breadth-first algorithm, which terminates when all terms are explored.

**Table 1** Hyponym-hypernym knowledge harvesting framework

1. Given:
   a DAP hyponym pattern $P_i = \{$*concept* such as *seed* and *$\}$
   a DAP$^{-}$1 hyponym pattern $P_c = \{$* such as $term_1$ and $term_2\}$
   a root concept *root*
   a term called *seed* for $P_i$
2. build a query using $P_i$
3. submit $P_i$ to Yahoo! or other search engine
4. extract terms occupying the * position
5. take terms from step 4. and go to step 2
6. repeat steps 2–5 until no new terms are found
7. rank terms by *outDegree*
8. all terms with *outDegree* > 0, build a query using $P_c$
9. submit $P_c$ to Yahoo! or other search engine
10. extract concepts (hypernyms) occupying the * position
11. rank concepts by *inDegree*
12. for ∀ terms with *inDegree* > 1, check subordination to the *root* with CPT
13. use concepts passing CPT from step 12. as temporary *root* and go to step 2
14. repeat steps 2–13 until the user desires

Although the DAP lexico-syntactic pattern has a very specific structure, we noticed that erroneous information can still be acquired due to part-of-speech tagging errors or flawed facts on the Web. Therefore, we need to filter out the erroneous terms from the true ones. For the purpose, we incorporate the harvested terms in a directed graph $G = (V, E)$, where each vertex $v \in V$ is a candidate term for the *semantic class* and each edge $(u, v) \in E$ indicates that the term $v$ is extracted from the term $u$. A term $u$ is ranked by $outDegree(u) = \frac{\sum_{\forall (u,v)\in E}(u,v)}{|V|-1}$, which represents all outgoing edges from $u$ normalized by the total number of nodes in the graph. In a very large corpus, like the Web, we assume that a correct term is the one that frequently discovers many different terms in the DAP pattern. In our illustrative example from Fig. 1, terms with high outDegree are *tiger, puma* among others.

## 5.2 Hypernym harvesting

In the hypernym extraction phase (i.e. extraction of concepts located above the low-level concepts of the taxonomy), we take all ⟨X, Y⟩ term pairs collected during the hyponym harvesting stage and instantiate them in the inverse DAP$^{-1}$ pattern *"* such as ⟨X⟩ and ⟨Y⟩"*. The pattern is sent to Yahoo! as a web query and all snippets matching the pattern are retrieved. For each ⟨X, Y⟩ pair, the terms discovered on the (*) position are extracted and considered as candidate hypernyms. For example, if the term *"cats"* was learned from the DAP pattern *"animals such as dogs and ⟨Y⟩"*, then the pair <*dogs,cats*> is used to form the new DAP$^{-1}$ query *"* such as dogs and cats"*, which extracts hypernyms such as *pets, mammals, others*.

To avoid the inclusion of erroneous hypernyms like *others*, we build a bipartite graph $G' = (V', E')$. The set of vertices $V_{sup}$ represents the hypernyms, while the set of vertices $V_p$ corresponds to the $\langle$X, Y$\rangle$ term pair that produced the hypernym. An edge $e'(u', v') \in E'$, where $u' \in V_p$ and $v' \in V_{sup}$ shows that the pair $\langle$X, Y$\rangle$ denoted as $u'$ harvested the hypernym represented by $v'$. Following the previous example, the bipartite graph would have three vertices $v'_1$, $v'_2$ and $v'_3$ for the hypernyms *"pets", "mammals", "others"*, one vertex $u'_1$ for the instance pair $\langle$*dogs, cats*$\rangle$, and three edges $e'_1(u'_1, v'_1)$, $e'_2(u'_1, v'_2)$ and $e'_3(u'_1, v'_3)$. A vertex $v' \in V_{sup}$ is ranked by $inDegree(v') = \frac{\sum_{\forall(u',v')\in E'}(u',v')}{|V'|-1}$, which represents the sum of all incoming edges to the hypernym node $v'$ from the term pairs $u'$. Intuitively, our confidence in a correct hypernym increases when it is discovered multiple times by different hyponym pairs.

## 5.3 Domain filtering

Although the aforementioned graph ranking functions can eliminate erroneous concepts, they cannot actually determine whether a concept is more or less general than the initial root concept. For example, when harvesting the categories (hypernyms) related to animals, the system may learn the word *species*, which is a very common term associated with animals, but also it applies to non-animal terms such as plants. To constrain the harvesting process to learn terms in a specific domain say *Animals*, we apply the *Concept Positioning Test (CPT)* that keeps only those terms that are located 'below' the initial root term. The CPT mechanism consists of two queries:

(a) *RootConcept such as Concept*
(b) *Concept such as RootConcept*

where *Concept* is the extracted hypernym and *RootConcept* is the starting root term. If the system returns more Web hits for (a) than (b), this indicates that the *Concept* passes the CPT test and it is located below the root. If the system returns more Web hits for (b) than (a) this means that the concept is more general than the root and it fails the CPT test and must be excluded from the domain.

To further augment the hyponym-hypernym term extractions of our knowledge harvesting algorithm, we use the concepts that pass the CPT test to build new DAP queries and then we re-instantiate the knowledge harvesting procedure from the very beginning. In this way we create a bootstrapping loop between the hyponym and hypernym phases. Note that this bootstrapping is separate from the local bootstrapping which is incorporated in the hyponym extraction phase. To instantiate the next hyponym-hypernym bootstrapping iteration, we replace the original root concept with the newly ranked hypernym and use all terms that lead to its discovery as seeds. Following our example, the hypernym *others* fails the CPT test because the term is more general than the root *animals*, while *pets* and *mammals* pass the criteria successfully. Next, we re-instantiate the original DAP pattern with two new patterns: "*pets such as dogs and* *" and "*mammals such as dogs and* *", where *pets* and *mammals* are the new *semantic class* concepts and *dogs* is the seed term as it discovered the hypernyms as shown in Sect. 5.2 The replacement of the initial root concept *animals* with the subordinated terms *pets* and *mammals* leads to the automated creation of new lexico-syntactic patterns that can extract terms which might have not been found with the initial DAP pattern "*animals such as* * *and* *". The described harvesting procedures in Sects. 5.1 and 5.2 can be repeated for unlimited number of iterations. For practical reasons we ran the algorithm for 10 iterations.

## 6 Taxonomy induction

The second stage of our algorithm concerns the hierarchical organization of the harvested knowledge. Next, we propose a graph-based algorithm, which positions the concepts with respect to each other and produces a taxonomy.

### 6.1 Positioning intermediate concepts

Once the knowledge acquisition and domain filtering phase terminates, we can obtain the is-a relations between the root and the low-level terms, as well as the is-a relations between the low-level and intermediate-level terms. However, the only information that is missing is the is-a relatedness of the intermediate-level concepts themselves.[2] For example, the knowledge harvesting algorithm does not provide information of the hierarchical organization of concepts like *mammals, carnivores, vertebrates, felines, chordates* among others.

Since the CPT test is an extremely reliable mechanism for the positioning of hypernyms with respect to the root, we decided to use the same procedure for the positioning the intermediate-level concepts. To gain more evidence from the Web, we use multiple surface patterns of the form: "X *such as* Y", "X *are* Y *that*", "X *including* Y", "X *like* Y", "*such* X *as* Y", where the *X* and *Y* corresponds to intermediate-level concepts. For instance, if we want to position the intermediate concepts *chordates* and *vertebrates* with respect to each other, we issue the CPT queries of the form: (a) *chordates such as vertebrates* and (b) *vertebrates such as chordates*. We record the counts of each pattern and estimate whether (a) returns more hits than (b). If this is the case, then *chordates* subsumes (or is broader than) *vertebrates*, otherwise *vertebrates* subsumes *chordates*.

### 6.2 Graph-based taxonomization

The left side of Fig. 2 visualizes the organization of the root, low-level and intermediate-level concepts according to the concept positioning mechanism. We can see that CPT cannot always determine the direct taxonomic organization between two concepts. For example, there is no is-a link between *felines* and *chordates* or between *felines* and *vertebrates*. One of the reasons is that these concepts are located on distant taxonomic levels and humans tend to exemplify

[2] The intermediate-level terms are located between the low-level and the root terms.

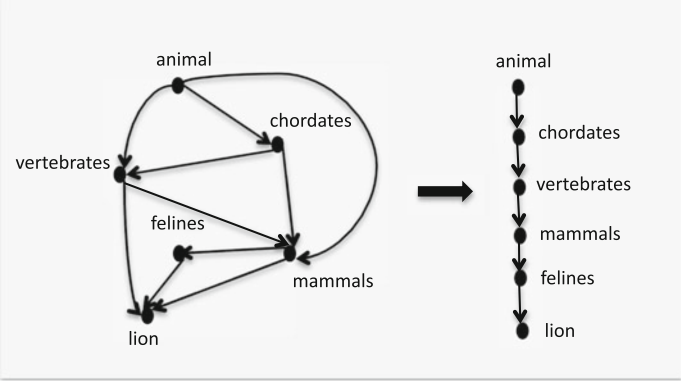

**Fig. 2** Concept positioning procedure and induced taxonomy

concepts using terms from proximate taxonomic levels. Therefore, CPT can find evidence for some is-a relations like *mammals* → *felines*, but not for others like *chordates* → *felines*.

After the concept positioning procedure has explored the positioning of all intermediate concept pairs, we observed two phenomena: (1) direct links between some concepts are missing and (2) multiple paths can be taken to reach from one concept to another.

To surmount these problems, we build a directed graph $G'' = (V'', E'')$ in which for a given a set of concepts (root, low, intermediate level ones), the objective is to find the longest path in the graph. In our case, the longest path would represent the taxonomic organization of the concepts as shown on the right side of Fig. 2.

In the graph $G''$, the nodes $V'' = \{t_1, t_2, t_3, \ldots, t_n, r\}$ represent the harvested terms (root, low, intermediate level), the edge $(t_i, t_j) \in E''$ indicates the is-a relatedness of $t_i$ and $t_j$, and the direction $t_i \rightarrow t_j$ corresponds to the term subordination according to the CPT test. If present, we eliminate all cycles in the graph. For that we use the CPT values of the terms and we use those whose weight is higher. If both terms have equal CPT values for (a) and (b), then we randomly select whether (a) or (b) subordination should remain. For each low-level term, we extract all hypernyms and is-a relations and use them to build a graph. On the top, we position the node with no predecessors $p$ (e.g. *animal*) and at the bottom, the node with no successor $s$ (e.g. terms like *lion*, *tiger*, *puma*). The directed graph is represented as an adjacency matrix $A = [a_{i,j}]$, where $a_{i,j}$ is 1 if $(t_i, t_j)$ is an edge of $G''$, and 0. To find the longest path between $p$ and $s$ pair, we find all possible paths between $p$ with $s$, and select the longest one among them.[3] We use this path to represent the taxonomic organization of all concepts located between $p$ and $s$. Once the taxonomization of a given low-level concept and its hypernyms terminates, we apply the same procedure to the next low-level term and its hypernyms.

[3] To compute the longest path we use a standard implementation.

## 7 Evaluation

### 7.1 Data collection and experimental set up

It is impossible to collect and report on results for all terms and domains. Therefore, to evaluate the effectiveness of our knowledge harvesting and taxonomization algorithm, we have selected the following four domains: *Animals, People, Vehicles*, and *Plants*. We choose these domains based on their diverse nature and characteristics, as well as the fact that they have taxonomic structures that are well-represented in WordNet.

We have instantiated the knowledge harvesting procedure with the following seed terms: *lions* for *Animals, Madonna* for *People, cars* for *Vehicles*, and *cucumbers* for *Plants*. To collect the data, we have submitted the DAP patterns as web queries to Yahoo!, retrieved the top 1,000 web snippets per query, and kept only the unique ones. In total, we have collected 10 GB of text snippets. We ran the hyponym extraction algorithm until complete exhaustion, while the hyponym-hypernym replacement steps for 10 iterations. The harvested data and the gold standard data used for our taxonomization evaluation can be downloaded here.[4]

At the end of the knowledge harvesting process, we found that the algorithm learned a staggering variety of terms, in far greater diversity than we had anticipated. In addition to many low-level terms, such as *dog, fox*, and *platypus*, and many intermediate terms, such as *predators, mammals, arachnids*, the algorithm has also harvested terms that are difficult to judge whether they are legitimate and valuable subconcepts of *Animals*. For instance, *bait, allergens, seafood, vectors, protein*, and *pests*. Another issue concerning the harvested concepts involves the relative terms that are hard to define in an absolute sense, such as *native animals* and *large mammals*.

Therefore, we believe that a complete evaluation of our task should answer the following three questions:

1. Precision: What is the correctness of the harvested concepts? (How many of them are simply wrong, given the root concept?)
2. Recall: What is the coverage of the harvested concepts? (How many are missing, below a given root concept?)
3. 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 something, 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

[4] http://www.isi.edu/~kozareva/data/kozareva_taxonomy_data.zip.

annotators judge as many results as we could afford (to obtain a measure of Precision and the legitimate extensions beyond WordNet).

In the next subsections we describe the obtained results for four different experiments conducted on the *Animals, People, Vehicles* and *Plants* domains. In Experiment 1, we evaluate the performance of DAP for hyponym learning, in Experiment 2, we evaluate the performance of DAP$^{-1}$ for hypernym learning, in Experiment 3, we evaluate the generated is-a relations between the concepts and in Experiment 4, we evaluate the induced taxonomic structures. For each experiment we conducted only a human-based evaluation and a comparative study against WordNet version 3.0. Initially, we also wanted to compare our results to knowledge bases that have been extracted in a similar way (i.e., through pattern application over unstructured text). However, it is not always possible to perform a complete comparison, because either researchers have not fully explored the same domains we have studied, or for those domains that overlap, the gold standard data was not available.

## 7.2 Experiment 1: hyponyms harvesting

In this section we discuss the results of the hyponym harvesting. The bootstrapping algorithm ranks the harvested terms by their *outDegree* score and considers as correct only those with *outDegree* > 0. In ten iterations, the bootstrapping algorithm produced 913 animal, 1,344 people, 1,262 plant and 1,425 vehicle terms that passed the *outDegree* criterion.

### *7.2.1 Human evaluation*

We employed two human judges to evaluate whether the harvested terms are correct or incorrect with respect to the root concept. Since human based evaluation for all harvested terms is time consuming and costly, we have evaluated all *Animals* and *People* terms, while for the *Vehicles* and *Plants* domains we have randomly selected 90 terms located at the beginning, in the middle and in the end of the *outDegree* ranking.

Figure 3 shows the Precision of the top *N* ranked terms. The overall performance of the *Animal* terms is 71 % (649/913) Precision and of the *People* terms is 95 % Precision (1,271/1,344). Figure 3 shows that higher-ranked *Animal* terms are more accurate than the lower-ranked terms, which indicates that the scoring function did its job. For *People* terms, precision was very high throughout the whole ranked list. The obtained results show that the hyponym step of the bootstrapping algorithm generates a large number of correct instances of high quality.

Table 2 summarizes the results for *Plants* and *Vehicles.*

Independently, we can say that the precision of the harvesting algorithm is between 73 and 90 % depending on the domains tested. In the case of *Vehicles*, we found that the learned terms in the middle ranking do not refer to the meaning of vehicle as a transportation device, but to the meaning of vehicle as media. Such extractions happen when both the class name and the term are ambiguous. For the

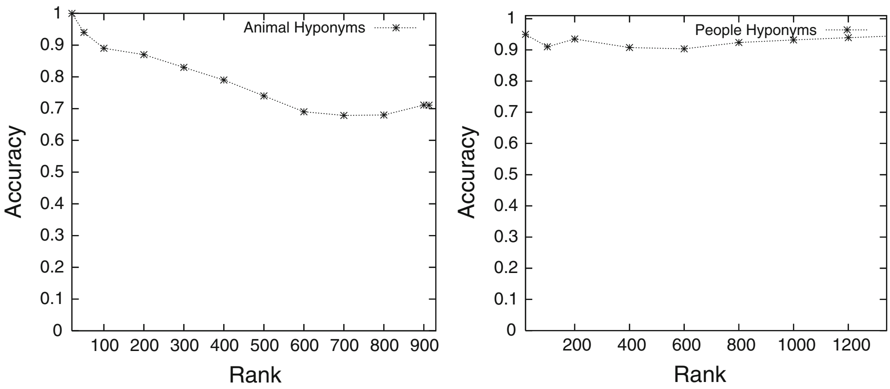

**Fig. 3** Ranked animals and people hyponyms

**Table 2** Human evaluation of plants and vehicles hyponyms

| | #CorrectByHand | #inWN | PrecByHand |
|---|---|---|---|
| Plants | | | |
| Rank (1–30) | 29 | 28 | .97 |
| Rank (420–450) | 29 | 21 | .97 |
| Rank (1,232–1,262) | 27 | 19 | .90 |
| Vehicles | | | |
| Rank (1–30) | 29 | 27 | .97 |
| Rank (193–223) | 22 | 18 | .73 |
| Rank (551–581) | 25 | 19 | .83 |

same category, the algorithm learned many terms which are not present in WordNet such as *BMW, bakkies, two-wheeler, all-terrain-vehicle* among others.

### *7.2.2 WordNet evaluation*

Table 3 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 and 986 correct *People* terms are not present in WordNet

**Table 3** WordNet hyponym evaluation

| | $Pr_{WN}$ | $Pr_H$ | *NotInWN* |
|---|---|---|---|
| Animal | .79 | .71 | 48 |
| People | .23 | .95 | 986 |

(primarily, for *People*, because WordNet contains relatively few proper names). These results show that there is substantial room for improvement in WordNet's coverage for these semantic classes. For *Animals*, the precision measured against WordNet is actually higher than the precision measured by human judges, which indicates that the judges failed to recognize some correct terms.

### *7.2.3 Evaluation against prior work*

As mentioned before, it is difficult to compare results with existing approaches, because either the researchers have not explored the same domains or for those domains that overlap the generated data is not available. Still to the extend to which it is possible, we compare the performance of our algorithm to the semantic class learning method of (Kozareva et al. 2008), which outperforms existing systems like those of (Pasca 2004) and KnowItAll (Etzioni et al. 2005).

The approach of (Kozareva et al. 2008) corresponds to the first step of our bootstrapping process. The difference between the current algorithm and those of (Kozareva et al. 2008) is in the hyponym-hypernym bootstrapping stage, which feeds on each iteration the newly learned intermediate-level concepts as roots for the DAP pattern and instantiates the learning from the very beginning.

We directly compare our results to (Kozareva et al. 2008), because the first iteration of our algorithm correspond to those of (Kozareva et al. 2008). Then, we ran the algorithm introduced in this paper for 10 hyponym-hypernym bootstrapping iterations and compared the obtained results. Figure 4 shows the number of harvested terms for *Animal* and *People* for each one of the 10 bootstrapping iterations.

Overall, the bootstrapping with intermediate concept substitution of the initial root term produced nearly 5 times as many low-level terms (hyponyms) compared to (Kozareva et al. 2008). It is important to note that not only the recall of the extractions was improved, but also the high precision of the extractions was maintained. Our observation is that the inclusion of the intermediate-level concepts in the hyponym extraction phase steered the learning process into new (yet still

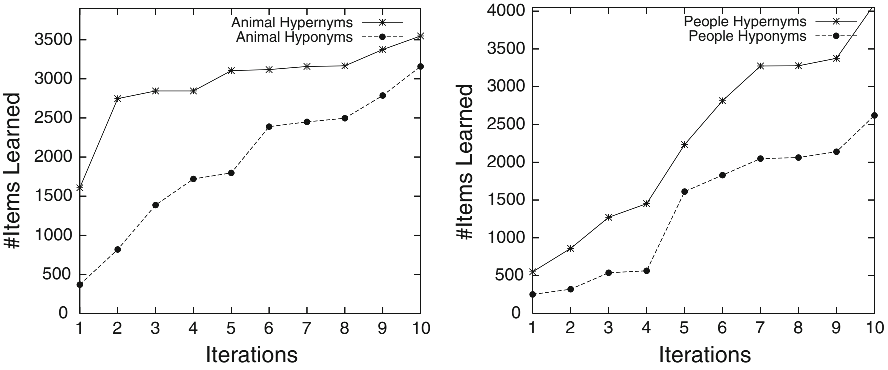

**Fig. 4** Learning curves for animals and people

**Table 4** Learned people and animals hyponym terms

| People | |
|---|---|
| Intermediate concept | 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 | |
|---|---|
| Intermediate concept | 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 |

correct) regions of the search space. For instance, the pattern "*animals such as * and **" covered parts of the *Animals* space, however the additionally generated hypernym patterns like "*herbivores such as * and **", "*mammals such as * and **" found previously unexplored parts of the *Animals* concept space.

Table 4 shows examples of the 10 top-ranked low-level terms that were learned for 3 randomly-selected intermediate-level *Animals* and *People* concepts (*Intermediate Concepts*) that were learned during bootstrapping.

## 7.3 Experiment 2: hypernym harvesting

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

### *7.3.1 Human evaluation and annotation guidelines*

We hired four annotators (undergraduates in linguistics from a different institution) to judge the correctness of the intermediate concepts. Unlike previous work on hypernym extraction (Ritter et al. 2009), where hypernyms were judged only as correct, incorrect, we created detailed annotation guidelines and categories which allow us to better understand the types and kinds of the harvested information. We defined 14 annotation labels for each one of the *Animals* and *People* classes. In the end, we cluster the fine-grained labels into four major types: *Correct, Borderline, BasicConcept*, and *NotConcept* to make it a consistent evaluation with respect to the rest of the hypernym approaches.

The annotation guidelines are as follows:

For our in-depth annotation, we have trained the undergraduate students in linguistics and asked them to classify each term as one or more of the types below. The annotators are instructed to: Try not to search for extreme and unusual interpretations of the term, but focus on the normal everyday meaning. Many terms have multiple senses. Here we are concerned only with the Animal/Human sense: if any of the senses is animal-related, then classify it based on that animal sense. Sometimes, a term might even refer to two different classes like cat (the house animal and the class, which includes tigers and lions). Thats fine; if you know of two senses that are not extreme or unusual cases, please code both (and make a Memo). Do this even if the snippets refer to only one of the classes. Please consider humans to be animals as well. That is, do not code GeneralTerm for a word like Hunter just because humans can be hunters too. Use GeneralTerm only if it includes things that are not humans or animals. The terms are expressed using a snippet of text derived from Google. These snippets dont necessarily correspond to complete sentences, nor do they typically respect sentence boundaries. You may use the snippet to understand whether the term refers to one of the codable categories, but dont be too concerned if one or more of the snippets isnt fully descriptive, representative, or even grammatical. When you dont know what a word means, or know enough to know which category(ies) it belongs to, you may use reference works to look up word meanings. If the snippets dont help, but you can ascertain the meaning of the word, you can go ahead and code it. If you really are stuck, but youre pretty sure the term refers to an animal, use code A (for OtherAnimal).

1. **BasicAnimal** The basic individual animal. Can be visualized mentally. Examples: Dog, Snake, Hummingbird.
2. **GeneticAnimalClass** A group of basic animals, defined by genetic similarity. Cannot be visualized as a specific type. Examples: Reptile, Mammal. Note that sometimes a genetic class is also characterized by distinctive behavior, and so should be coded twice, as in Sea-mammal being both GeneticAnimalClass and BehavioralByHabitat. (Since genetic identity is so often expressed as body structureits a rare case that two genetically distant things look the same structurallyit will be easy to confuse this class with MorphologicalTypeAnimal. If the term refers to just a portion of the animal, its probably a MorphologicalTypeAnimal. If you really see the meaning of the term as both genetic and structural, please code both.)
3. **NonRealAnimal** Imaginary animals. Examples: Dragon, Unicorn. (Does not include normal animals in literature or films.)
4. **BehavioralByFeeding** A type of animal whose essential defining characteristic relates to a feeding pattern (either feeding itself, as for Predator or Grazer, or of another feeding on it, as for Prey). Cannot be visualized as an individual animal. Note that since a term like Hunter can refer to a human as well as an animal, it should not be classified as GeneralTerm.

5. **BehavioralByHabitat** A type of animal whose essential defining characteristic relates to its habitual or otherwise noteworthy spatial location. Cannot be visualized as an individual animal. (When a basic type also is characterized by its spatial home, as in South African gazelle, treat it just as a type of gazelle, i.e., a BasicAnimal. But a class, like South African mammals, belongs here.) Examples: Saltwater mammal, Desert animal. And since a creatures structure is sometimes determined by its habitat, animals can appear as both; for example, South African ruminant is both a BehavioralByHabitat and a MorphologicalTypeAnimal.
6. **BehavioralBySocializationIndividual** A type of animal whose essential defining characteristic relates to its patterns of interaction with other animals, of the same or a different kind. Excludes patterns of feeding. May be visualized as an individual animal. Examples: Herding animal, Lone wolf. (Note that most animals have some characteristic behavior pattern. So use this category only if the term explicitly focuses on behavior.)
7. **BehavioralBySocializationGroup** A natural group of basic animals, defined by interaction with other animals. Cannot be visualized as an individual animal. Examples: Herd, Pack.
8. **MorphologicalTypeAnimal** A type of animal whose essential defining characteristic relates to its internal or external physical structure or appearance. Cannot be visualized as an individual animal. (When a basic type also is characterized by its structure, as in Duck-billed platypus, treat it just as a type of platypus, i.e., a BasicAnimal. But a class, like Armored dinosaurs, belongs here.) Examples: Cloven-hoofed animal, Short-hair breed. And since a creatures structure is sometimes determined by its habitat, animals can appear as both; for example, South African ruminant is both a MorphologicalType-Animal and a BehavioralByHabitat. Finally, since genetic identity is so often expressed as structureits a rare case that two genetically distant things look the same structurallyit will be easy to confuse this class with MorphologicalType-Animal. If the term refers to just a portion of the animal, its probably a MorphologicalTypeAnimal. But if you really see both meanings, genetic and structural, please code both.
9. **RoleOrFunctionOfAnimal** A type of animal whose essential defining characteristic relates to the role or function it plays with respect to others, typically humans. Cannot be visualized as an individual animal. Examples: Zoo animal, Pet, Parasite, Host.

G. **GeneralTerm** A term that includes animals (or humans) but refers also to things that are neither animal nor human. Typically either a very general word such as Individual or Living being, or a general role or function such as Model or Catalyst. Note that in rare cases a term that refers mostly to animals also includes something else, such as the Venus Fly Trap plant, which is a carnivore. Please ignore such exceptional cases. But when a large proportion of the instances of a class are non-animal, then code it as GeneralTerm.

E. **EvaluativeTerm** A term for an animal that carries an opinion judgment, such as varmint. Sometimes a term has two senses, one of which is just the animal, and the other is a human plus a connotation. For example, snake or weasel is either the animal proper or a human who is sneaky; lamb the animal proper or a person who is gentle, etc. Since the term can potentially carry a judgment connotation, please code it here as well as wherever else the animal sense of it belongs.

A. **OtherAnimal** Almost certainly an animal or human, but none of the above applies, or: I simply dont know enough about the animal to know where to classify it.

0. **NotAnimal** Not an animal or human. But a real English term nonetheless.

B. **GarbageTerm** Not a real English word.

For People we have defined the following categories.

1. **BasicPerson** The basic individual person or persons. Can be visualized mentally. Examples: Child, Woman.
2. **GeneticPersonClass** A person or persons defined by genetic charactertics/similarity. Can be visualized as a specific type. Examples: Asian, Saxon. Note that sometimes a genetic class is also characterized by nationality or tribal affiliation, and so should be coded twice, as in Eskimo.
3. **ImaginaryPeople** Imaginary individuals or groups. Examples: Superman, the Hobbits. human-like creatures such as elves and dwarves, as well as normal people in literature or films, such as Tom Sawyer.
4. **RealPeople** Specific real individuals or groups, by name or description. Example: Madonna, Mother Theresa, the Beatles, the first man on the moon, Marco Polo, the person who invented the wheel.
5. **NonTransientEventParticipant** The role a person plays consistently over time, by taking part in one or more specific well-defined events. Sometimes, a word may be ambiguous between an ongoing/repeated event and a transient one; please code both (examples: donor, as someone who tends to give, or who only gives once; well-wisher; mentor). Distinguishing this class from PersonState, there is always an associated characteristic action or activity that either persists or recurs, without a specific endpoint being defined. This group includes several types, including: Occupations (priest, doctor), Hobbies (skier, collector), Habits (stutterer, peacemaker, gourmand).
6. **TransientEventParticipant** The role a person plays for a limited time, through taking part in one or more specific well-defined events. There is always an associated characteristic action or activity, with a defined (though possibly unknown) endpoint. The duration of the event is typically from hours to days, perhaps up to a year, but certainly less than a decade. Examples: speaker, passenger, visitor. If the role lasts longer (say, a rivalry over years), then use PersonState. Sometimes, a word may be ambiguous between a transient event and an ongoing/repeated one; please code both (examples: donor, as someone who tends to give, or who only gives once; well-wisher; mentor).
7. **PersonState** A person with a certain physical or mental characteristic that persists over time. Distinguishing this class from NonTransientEventParticipant, there is no typical associated defining action or activity that one can

think of. Examples: midget, schizophrenic, AIDS patient, blind person. (One could distinguish subtypes of PersonState—say PersonStateMental and PersonStatePhysical—which would neatly place Schizophrenic and Liberal (! together) in the former, and BlindPerson and Midget in the latter.) Note that PersonState is neither a social role nor a NationOrTribal one, so it does not include socialite, being a mother, or being Japanese.

8. **FamilyRelation** A family relation. Examples: aunt, mother. This is a specialized subcategory of SocialRole, so dont code family relations twice.
9. **SocialRole** The role a person plays in society. Unlike NonTransientEventParticipant, there is no single associated defining event or activity, but rather a collection of possible ones together. (Even professions that may involve many different activities, such as president and secretary, and family relations, such as mother, do not belong here.) Always, however, the role relates to other people in some clear social setting. Examples: role model, fugitive, alumnus, hero, star, guest. The intention is that SocialRole captures notions like Leader (in its general sense), since it's not associated with any single clearly defined action. NonTransientEventParticipants like President, Boss, or Leader (in its narrow sense, as Patrol Leader), all have several specific duties to fulfill, many of which make them be leaders (in the general sense).

N. **NationOrTribe** A nationality or tribal affiliation. Examples: Bulgarian, American, Swiss, Zulu. Note that aboriginal is a GeneticPersonClass, not a NationOrTribe.

R. **ReligiousAffiliation** A religious affiliation. Examples: Catholic, atheist. Some religious affiliations, notably being Jewish, have strong NationOrTribe connotations as well; please code both.

H. **OtherHuman** Clearly a human and not an animal or other being, but does not fit into any other class.

G. **GeneralTerm** Can be a human, but also includes other non-human entities. Examples: image, example, figure.

0. **NotPerson** Simply not a person.

More information on the detailed annotation guidelines and the annotation study can be found in (Hovy et al. 2009).

Table 5 summarizes the labels we have defined as well as examples of some terms corresponding to each category. We measured the pairwise inter-annotator agreement across the fourteen labels using the Fleiss kappa (Fleiss 1971). The κ scores ranged from 0.61–0.71 for *Animals* (average $\kappa = 0.66$) and from 0.51–0.70 for *People* (average $\kappa = 0.60$). These agreement scores seemed good enough to warrant the usage of these human judgments to estimate the precision of the algorithm, however they also showed that the task is not trivial.

In ten iterations, the bootstrapping algorithm harvested 3,549 *Animal* and 4,094 *People* intermediate-level concepts. After the *inDegree* ranking was applied, we selected a random sample of intermediate-level concepts and gave them for annotation to the four human judges. Table 6 shows the labels assigned by the four annotators ($A_1 - A_4$).

**Table 5** Intermediate concept annotation labels

| Type | Label | Examples |
|---|---|---|
| *Animal* | | |
| 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* | | |
| 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* |

The top portion of Table 6 shows the results for all intermediate concepts (437 animal terms and 296 people terms), and the bottom portion shows the results only for those that passed the CPT (187 *Animal* terms and 139 *People terms*).

We compute the precision of the extracted terms 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 ranking the precision ranges from 53 to 66 % for *Animals* and 75–85 % for *People*. After applying the CPT ranking the precision increased to 71–84 % for *Animals* and 82–94 % for *People*. These results confirm that the CPT is effective at removing undesirable general terms. Overall, the results demonstrate that our algorithm produced many high-quality intermediate concepts, with good precision.

**Table 6** Human intermediate concept evaluation

| | 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 |

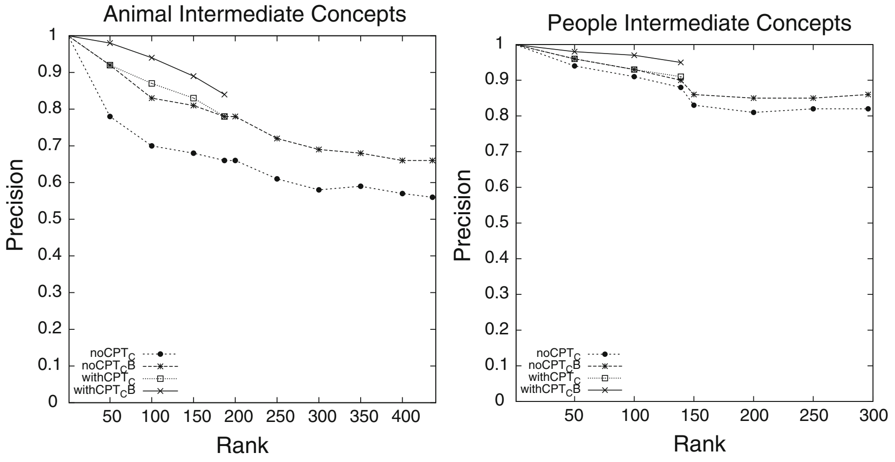

**Fig. 5** Precision of intermediate concepts at rank N

Next, we show in Fig. 5 the precision curves of the *inDegree* rankings of the intermediate concepts tagged as correct (*c*) or correct and borderline (*cb*) with the CPT (*withCPT*) and without the CPT (*noCPT*) test. Clearly, CPT improves the precision 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. In addition, CPT also eliminated many general terms. For *People* a much larger set of intermediate concepts was learned. Precision remained relatively high even without the CPT test.

### *7.3.2 WordNet evaluation*

We also compared the precision and the presence of the automatically learned intermediate concepts against those found in WordNet. The obtained results are shown in Table 7.

Of the learned intermediate-level concepts, WordNet contains 20 % of the *Animals* and 51 % of the *People* terms. This confirms that many of the concepts were also considered to be valuable taxonomic terms by the WordNet developers. However, our human annotators found 66 % of the *Animals* and 85 % of the *People* concepts to be correct, which suggests that the algorithm generated a substantial amount of additional concepts that could be used to further enrich the taxonomic structure in WordNet.

## 7.4 Experiment 3: taxonomic links

In this section, we evaluate the learned is-a links between the harvested concepts. 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.

### *7.4.1 Human and WordNet evaluations*

Table 8 shows the results for the is-a relations between all terms (intermediate and low-level ones). For each pair, we extracted the harvested links and determined whether the same links appear in WordNet. We also gave the same links to human judges.

The results show that the DAP patterns can accurately extract is-a relations. It is important to note that a substantial portion of these relations is not present in WordNet. For example, from the manually evaluated relations, there are 804 *Animal* and 539 *People* links that are missing from WordNet.

**Table 7** WordNet intermediate concept evaluation

| | $Pr_{WN}$ | $Pr_H$ | NotInWN |
|---|---|---|---|
| Animal | .20 (88/437) | .66 (288/437) | 204 |
| People | .51 (152/296) | .85 (251/296) | 108 |

**Table 8** WordNet taxonomic evaluation

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

## 7.5 Experiment 4: reconstructing WordNet's taxonomy

In the final experiment, we evaluate the performance of our algorithm to induce a taxonomic structure for the concepts learned in a given domain. Since the manual construction and the evaluation of the harvested taxonomies is extremely challenging and difficult even for human experts, we decided to evaluate the performance of our algorithm only by reconstructing WordNet's *Animals, Plants* and *Vehicles* taxonomies. We did not evaluate the taxonomy for *People*, because most of the learned instances and hypernyms are missing from WordNet.

For each domain we selected the terms which were harvested by our algorithm and also present in WordNet. For each term and root concept (*Animal, Plant* or *Vehicle*) we retrieved all concepts located on the path between the two terms and used this information to evaluate our approach. Practically being able to reconstruct WordNet's taxonomy for these concepts is equivalent to evaluating the performance of our taxonomy induction approach.

Table 9 summarizes the characteristics of the taxonomies for the regions tested. For each domain, we show the total number of terms that must be organized, and the total number of is-a relations that must be induced.

Among the three domains we have used for our evaluation, the *Animals* one is the most complex and has the richest taxonomic structure. The maximum number of levels that must be inferred is 11, the minimum is 1 and the average taxonomic depth is 6.2. In total there are three low-level concepts (*longhorns, gaur* and *bullock*) with maximum depth, twenty terms (low-level and intermediate concepts) with minimum depth and ninety-eight low-level terms (*wombat, viper, rat, limpkin*) with depth 6. *Plants* is also a very challenging domain, because it contains a mixture of scientific and general terms such as *magnoliopsida* and *flowering plant*.

### 7.5.1 Taxonomy evaluation

To evaluate the performance of our taxonomy induction approach, we use the following measures:

$$Precision = \frac{\#\, is - a\, found\, in\, WordNet\, and\, by\, system}{\#\, is - a\, found\, by\, system}$$

$$Recall = \frac{\#\, is - a\, found\, in\, WordNet\, and\, by\, system}{\# is - a\, found\, in\, WordNet}$$

**Table 9** Data for WordNet reconstruction

| | Animals | Plants | Vehicles |
|---|---|---|---|
| #Terms | 684 | 554 | 140 |
| #Is-a | 4,327 | 2,294 | 412 |
| Average depth | 6.23 | 4.12 | 3.91 |
| Max depth | 12 | 8 | 7 |
| Min depth | 1 | 1 | 1 |

**Table 10** Evaluation of the induced vehicle taxonomy

| Vehicles | Precision | Recall |
|---|---|---|
| X such as Y | .99 (174/175) | .42 (174/410) |
| X are Y that | .99 (206/208) | .50 (206/410) |
| X including Y | .96 (165/171) | .40 (165/410) |
| X like Y | .96 (137/142) | .33 (137/410) |
| Such X as Y | .98 (44/45) | .11 (44/410) |
| All patterns | .99 (246/249) | .60 (246/410) |

**Table 11** Evaluation of the induced taxonomies

| | Precision | Recall |
|---|---|---|
| Animals | .98 (1,643/1,688) | .38 (1,643/4,327) |
| Plants | .97 (905/931) | .39 (905/2294) |
| Vehicles | .99 (246/249) | .60 (246/ 410) |

Table 10 shows results for the taxonomy induction of the *Vehicles* domain using different concept positioning patterns. The most productive patterns are: "X *are* Y *that*" and "X *including* Y", however the highest yield is obtained when we combine the evidence from all patterns (i.e. when we sum the retrieved Web counts from all patterns).

Table 11 shows results for the taxonomization of the *Animals, Plants*, and *Vehicles* domains.

Figure 6 shows an example of our taxonomy induction algorithm for some low-level terms like *vipers, rats, wombats, ducks, emus, moths*, and *penguins* and their hypernyms.

The obtained results are very encouraging given the fact that we started the taxonomy construction entirely from scratch (i.e. without the usage of a skeleton structure of any existing taxonomy). The precision of the taxonomization approach is very robust. However, recall must be further improved since not all concepts were found with the lexico-syntactic patterns.

Still the biggest challenge for any taxonomization approach is the merging of the independent taxonomic perspectives (a *deer* is a *grazer* in BehaviorByFeeding, a *wildlife* in BehaviorByHabitat, a *herd* in BehaviorSocialGroup and an *even-toed ungulate* in MorphologicalType) into a single hierarchy.

### *7.5.2 Comparative study on taxonomy evaluation*

Finally, we compare the performance of our pattern-based taxonomy induction algorithm with another contemporary graph-based taxonomization algorithm developed by (Navigli et al. 2011). Since they have used all of our harvested terms, is-a relations and gold standard data to evaluate the performance of their taxonomization algorithm, this is making it easy for us to conduct comparative studies and hopefully it would also encourage other researchers working on

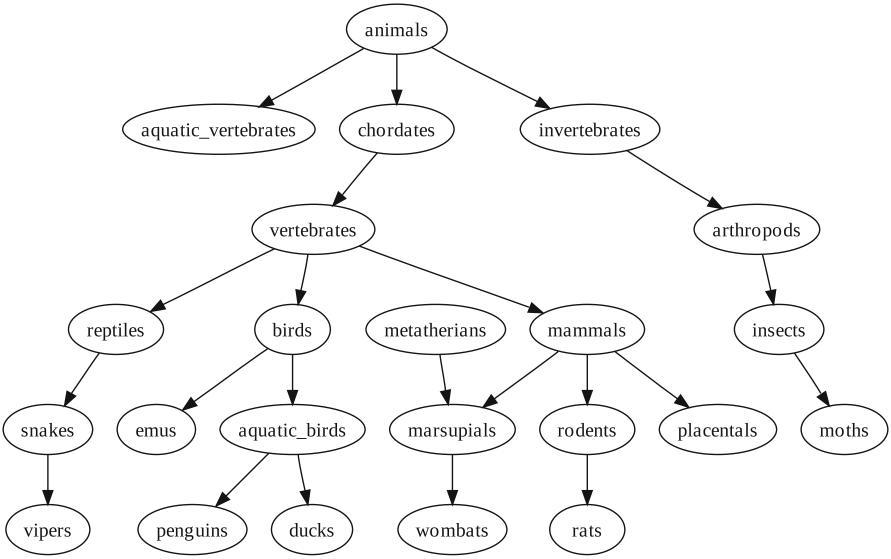

**Fig. 6** An example of the induced taxonomy of our algorithm for some animal terms

**Table 12** Comparative evaluation of our taxonomy induction algorithm and the graph-based taxonomy induction algorithm of Navigli et al. (2011)

| | Our approach | | Navigli et al. (2011) | |
|---|---|---|---|---|
| | Precision | Recall | Precision | Recall |
| Animals | .98 (1,643/1,688) | .38 (1,643/4,327) | .97 (1,638/1,688) | **.44 (1,890/4,327)** |
| Plants | .97 (905/931) | **.39 (905/2,294)** | .97 (905/931) | .38 (879/2,294) |
| Vehicles | **.99 (246/249)** | **.60 (246/410)** | .91 (226/249) | .49 (200/410) |

taxonomy induction to use our knowledge harvested data as a reference point for comparison.

To briefly summarize, our algorithm used CPT to find term relatedness, while (Navigli et al. 2011) used graph trimming and edge weighting procedure. In our case, we induce the taxonomy using the longest path in the graph, while (Navigli et al. 2011) used a Chu-Liu/Edmonds algorithm to find the optimal branching and then they applied pruning recovery to induce the final taxonomy.

Table 12 shows the obtained results of the two algorithms for the same number of terms, is-a relations and taxonomies. Our pattern-based taxonomy induction outperforms (Navigli et al. 2011) for two out of the three domains. We obtained lower recall only for the *Animals* domain. If we had the output of Navigli's system, we could analyze the obtained results to better understand what type information was missed by our algorithm, but unfortunately such information is not present.

In conclusion, we can say that the beauty of our work lies not only in the simplicity of our knowledge harvesting and taxonomization algorithm, which is making it easy to implement and use by anyone, but also in our effort to create and freely distribute a taxonomization data set, which can be used as an evaluation benchmark by other unsupervised taxonomy induction algorithms.

## 8 Discussion

It is clear that text harvesting can significantly assist with the creation of wordnets and ontologies. Finding all the terms in a given domain automatically greatly reduces the manual dictionary and wordlist search. But such harvesting also poses challenges: It is unlikely, for example, that a human wordnet builder would come up with the term *even-toed ungulate*. The hypernyms harvested as per Sect. 5.2 illustrate clearly that simple term taxonomies such as found in current wordnets and most ontologies are completely inadequate, and that some sort of multiple parallel taxonomization, such as discussed above, is required. Which kinds of parallel hierarchies are needed for which root concepts, however, is unclear. We believe that a start can be made with the observation that, for Entities, there appears to be three families of characteristics:

- **Structure**: This dimension of description includes *material* properties such as the materials that the entity is made of, *morphological* properties such as the parts of entities, and *articulatory* properties such as the ways in which the parts are assembled and connected.
- **Function**: This dimension includes the *purposes* of entities (why they were constructed, if they are artifacts), and the *applications* of entities, such as the manner in which people employ them to achieve the purposes.
- **Provenance**: This dimension includes various kinds of *sources* of the entities, including who built them and where they are made, grown, or found, as well as the *history* of the entity.

However, the equivalent conceptual breakdown for Events and States is much less apparent. These distinctions also map with findings on qualia structures by (Pustejovsky 1995) and (Moravcsik 1981) interpretation of Aristotle's modes of explanations.

The CPT taxonomization procedure described in Sect. 6 is a start, but works far better for some concepts than others. Events, states, and relations, and even complex AbstractEntities such as Emotions or InformationObjects such as *stories, symphonies, news*, etc., are very difficult even for humans to taxonomize. It may be the case that one can extend the CPT to obtain suggested folk taxonomizations directly from the harvested corpus; then whatever the 'truth' might be, one at least can fall back onto how the majority of authors in the corpus view the matter. A good example is the popular treatment of a *dolphin* as a Fish, even though it is biologically a Mammal.[5] Recent interest in folksonomies (Peters 2009) reflects the potential of this approach.

[5] The various approaches to such ontological decisions are discussed in Hovy (2002).

Evaluation remains a difficult matter. The terms harvested by DAP, even for such relatively well-researched concept families as the Animal Kingdom, sometimes far exceed the terms included in wordnets, making both Precision and Recall very expensive to measure. We need other methods to validate terms harvested by DAP and similar algorithms, for example using their distributional semantic properties.

Finally, we have not in this article discussed the use of DAP-like algorithms to harvest the properties of concepts (for example, that bees are small and pomegranates are red). But it is a direct extension of the basic DAP pattern to do so. Automatically constructing rich entity descriptions using this approach is an interesting challenge for the future.

## 9 Conclusion

In this article we demonstrate the effectiveness of a very simple class of text harvesting patterns, the recursive family we call DAP, to collect and partially taxonomize sets of terms conceptually subordinate to a given starting concept. We illustrate the power of DAP on a variety of starting concepts, and show how English WordNet, one of the largest and most complete online term taxonomies ever created, is still far from complete, when compared to language on the web. We show the need for more carefully considered taxonomization than has heretofore been the case in most taxonomies.

The construction of online wordnets in various languages is an important endeavor. We believe that by employing such algorithms as DAP and its subsidiary CPT, which individually assist with steps in the overall process rather than trying to achieve the whole ontology learning procedure at once, the work can be facilitated. There is still a long way to go, and a lot of fascinating research to be done.

**Acknowledgments** We acknowledge the support of DARPA contract number FA8750-09-C-3705.

## References

Agirre, E., & Lopez de Lacalle, O. (2004). Publicly available topic signatures for all WordNet nominal senses. In *Proceedings of the 4rd international conference on languages resources and evaluations* (LREC). Lisbon, Portugal.

Amsler, R. A. (1981). A taxonomy for english nouns and verbs. In: *Proceedings of the 19th annual meeting on association for computational linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, pp. 133–138.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., et al. (2004). The MEANING multilingual central repository. In *Proceedings of the second international WordNet conference*. pp. 80–210.

Banko, M. (2009). *Open information extraction from the web*. Ph.D. Dissertation from University of Washington.

Bateman, J. A., Kasper, R. T., Moore, J. D., & Whitney, R. A. (1989). *A general organization of knowledge for natural language processing: The penman upper model*. Unpublished research report, USC/Information Sciences Institute, Marina del Rey.

Cuadros, M., & Rigau, G. (2008). KnowNet: Building a large net of knowledge from the web. *The 22nd international conference on computational linguistics* (Coling'08), UK, Manchester.

Davidov, D., & Rappoport, A. (2006). Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *Proceedings of the 21st international conference on computational linguistics COLING and the 44th annual meeting of the ACL*, pp. 297–304.

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

Fellbaum, C. (Ed.). (1998). WordNet: An on-line lexical database and some of its applications. Cambridge, MA, MIT Press.

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

George, A. M. (1995). WordNet: A lexical database for english. *Proceedings of Communications of the ACM, 38* pp. 39–41.

Girju, R., Badulescu, A., & Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the conference of the north american chapter of the association for computational linguistics on human language technology* (NAACL-HLT), pp. 1–8.

Glickman, O., Dagan, I., & Koppel, M. (2005). A probabilistic classification approach for lexical textual entailment. In *Proceedings of the twentieth national conference on artificial intelligence and the seventeenth innovative applications of artificial intelligence conference*, pp. 1050–1055.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on computational linguistics*, pp. 539–545.

Hovy, E. H. (1998). Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In *Proceedings of the LREC conference*.

Hovy, E. H. (2002). Comparing sets of semantic relations in ontologies. In R. Green, C. A. Bean, & S. H. Myaeng (Eds.), *The semantics of relationships: An interdisciplinary perspective*, pp. 91–110.

Hovy, E. H., Kozareva, Z., & Riloff, E. (2009). Toward completeness in concept extraction and classification. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (EMNLP), pp. 948–957.

Hovy, E. H., & Nirenburg, S. (1992). Approximating an interlingua in a principled way. In *Proceedings of the DARPA Speech and natural language workshop*, Arden House, NY.

Ide, N., & Veronis, J. (1994). Machine readable dictionaries: What have we learned, where do we go. In *Proceedings of the post-COLING 94 intl. workshop on directions of lexical research*, Beijing, pp. 137–146.

Katz, B., & Lin, J. (2003). Selectively using relations to improve precision in question answering. In *Proceedings of the EACL-2003 workshop on natural language processing for question answering*, pp. 43–50.

Kozareva, Z., Riloff, E., & Hovy, E. H. (2008). Semantic class learning from the web with hyponym pattern linkage graphs. In *Proceedings of the NAACL-HLT conference*, pp. 1048–1056.

Lenat, D. B., & Guha, R. V. (1990). *Building large knowledge-based systems. reading*. Boston: Addison-Wesley.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on computational linguistics* (COLING), pp. 768–774.

Lin, D., & Pantel, P. (2002). Concept discovery from text. In *Proceedings of the 19th international conference on computational linguistics* (COLING), pp. 1–7.

Miller, G. A. (1995). WordNet: a lexical database for english. *Communications of the ACM, 38*, 39–41.

Mitchell, T. M., Betteridge, J., Carlson, A., Hruschka, E., & Wang, R. (2009). Populating the semantic web by macro-reading internet text. In *Proceedings of the 8th international semantic web conference* (ISWC).

Moldovan, D. I., Harabagiu, S. M., Pasca, M., Mihalcea, R., Goodrum, R., Girju, R. et al. (1999). Lasso: A tool for surfing the answer net. In *Proceedings of the TREC conference*.

Moravcsik, J. M. E. (1981). How do words get their meanings? *The Journal of Philosophy, 78* 1.

Navigli, R., & Ponzetto, P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Journal of Artificial Intelligence, 193*, 217–250.

Navigli, R., Velardi, P., Cucchiarelli, A., Neri, F., & Cucchiarelli, R. (2004). Extending and enriching WordNet with OntoLearn. In *Proceedings of the second global wordnet conference 2004* (GWC 2004). pp. 279–284.

Navigli, R., Velardi, P., & Faralli, S. (2011). A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of the Twenty-Second international joint conference on artificial intelligence—volume volume three*. IJCAI'11, pp. 1872–1877.

Pantel, P., & Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of 21st international conference on computational linguistics (COLING) and 44th annual meeting of the association for computational linguistics (ACL)*.

Pantel, P., Crestan, E., Borkovsky, A., Popescu, A. M., & Vyas, V. (2009). Web-scale distributional similarity and entity set expansion. In *Proceedings of the conference on empirical methods in natural language processing* (EMNLP), pp. 938–947.

Pasca, M. (2004). Acquisition of categorized named entities for web search. In *Proceedings of the thirteenth ACM international conference on information and knowledge management* (CIKM), pp 137–145.

Pease, A., Fellbaum, C., & Vossen, P. (2008). Building the global WordNet grid. In *Proceedings of the 18th international congress of linguists* (CIL18), Seoul, Republic of Korea, July, pp. 21–26.

Pennacchiotti, M., & Pantel P. (2006). Ontologizing semantic relations. In *Proceedings of the international conference on computational linguistics (COLING) and the annual meeting of the association for computational linguistics* (ACL), pp. 793–800.

Peters, I. (2009). *Folksonomies. Indexing and retrieval in web 2.0*. Berlin: De Gruyter Saur.

Ponzetto, S., & Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (ACL 2010), Uppsala, Sweden.

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.

Richardson, S. D., Dolan, W. B., & Vanderwende, L. (1998). Mindnet: Acquiring and structuring semantic information from text. In *Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics*—Volume 2 (ACL '98), (Vol. 2). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1098–1102.

Rigau, G., Rodriguez, H., & Agirre, E. (1998). Building accurate semantic taxonomies from monolingual MRDs. In *Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics*—Volume 2 (ACL '98), (Vol. 2). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1103–1109.

Riloff, E., & Shepherd, J. (1997). A corpus-based approach for building semantic lexicons. In *Proceedings of the second conference on empirical methods in natural language processing* (EMNLP), pp. 117–124.

Riloff, E., & Jones, R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the sixteenth national conference on artificial intelligence* (AAAI), pp. 474–479.

Ritter, A., Soderland, S., & Etzioni, O., (2009). What is this, anyway: Automatic hypernym discovery. In *Proceedings of the AAAI spring symposium on learning by reading and learning to read*.

Ritter, A., & Mausam, O.E. (2010). A latent dirichlet allocation method for selectional preferences. In *Proceedings of the association for computational linguistics conference* (ACL).

Roberto, N., Velardi, P., & Faralli, S. (2011). A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of IJCAI 2011*, pp. 1872–1877.

Robkop, K., Thoongsup, S., Charoenporn, T., Sornlertlamvanich, V., & Isahara, H. (2010). *WNMS: Connecting the distributed WordNet in the case of Asian WordNet the 5th international conference of the global WordNet association* (GWC-2010), Mumbai, India.

Rosch, E. (1978). Principles of categorization. In *Cognition and Categorization*, pp. 27–48

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, pp. 4449.

Snow, R., Jurafsky, D., & Ng, A.Y. (2005). Learning syntactic patterns for automatic hypernym discovery. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 17, pp. 1297–1304).

Snow, R., Jurafsky, D., & Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the international conference on computational linguistics (COLING) and the annual meeting of the association for computational linguistics* (ACL).

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: A core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web* (WWW), pp. 697–706.

Szpektor, I., Dagan, I., Bar-Haim, R., & Goldberger, J. (2008). Contextual preferences. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)*, pp. 683–691.

Velardi, P., Roberto, N., & Pierluigi, D. (2008). Mining the web to create specialized glossaries. *Journal of IEEE Intelligent Systems, 23*(5) 18–25. ISSN:1541-1672.

Vossen, P., Hofmann, K., Rijke, M., Tjong, E., Sang, K., & Deschacht, K. (2008). The Cornetto database: Architecture and user-scenarios. In *Proceedings of the fourth international GlobalWordNet conference—GWC*.

Vossen, P. (Ed.). (1998). *EuroWordNet: A multilingual database with lexical semantic networks*. Dordrecht, The Netherlands: Kluwer.

Widdows, D. (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of the HLT-NAACL conference*.

Wilks, Y., Fass, D., ming Guo, C., Mcdonald, J. E., Plate, T., & Slator, B. M. (1988). Machine tractable dictionaries as tools and resources for natural language processing. In *Proceedings of the 12th conference on computational linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, pp. 750–755.

Yang, H., & Callan, J. (2009). A metric-based framework for automatic taxonomy induction. In *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP* (ACL-IJCNLP) (Vol. 1, pp. 271–279.