Semantic Taxonomy Induction from Heterogenous Evidence

Rion Snow
Computer Science Department
Stanford University
Stanford, CA 94305
rion@cs.stanford.edu

Daniel Jurafsky
Linguistics Department
Stanford University
Stanford, CA 94305
jurafsky@stanford.edu

Andrew Y. Ng
Computer Science Department
Stanford University
Stanford, CA 94305
ang@cs.stanford.edu

Abstract

We propose a novel algorithm for inducing semantic taxonomies. Previous algorithms for taxonomy induction have typically focused on independent classifiers for discovering new single relationships based on hand-constructed or automatically discovered textual patterns. By contrast, our algorithm flexibly incorporates evidence from multiple classifiers over heterogenous relationships to optimize the entire structure of the taxonomy, using knowledge of a word's coordinate terms to help in determining its hypernyms, and vice versa. We apply our algorithm on the problem of sense-disambiguated noun hyponym acquisition, where we combine the predictions of hypernym and coordinate term classifiers with the knowledge in a preexisting semantic taxonomy (WordNet 2.1). We add $10,000$ novel synsets to WordNet 2.1 at 84% precision, a relative error reduction of 70% over a non-joint algorithm using the same component classifiers. Finally, we show that a taxonomy built using our algorithm shows a 23% relative F-score improvement over WordNet 2.1 on an independent testset of hypernym pairs.

1 Introduction

The goal of capturing structured relational knowledge about lexical terms has been the motivating force underlying many projects in lexical acquisition, information extraction, and the construction of semantic taxonomies. Broad-coverage semantic taxonomies such as WordNet (Fellbaum, 1998) and CYC (Lenat, 1995) have been constructed by hand at great cost; while a crucial source of knowledge about the relations between words, these taxonomies still suffer from sparse coverage.

Many algorithms with the potential for automatically extending lexical resources have been proposed, including work in lexical acquisition (Riloff and Shepherd, 1997; Roark and Charniak, 1998) and in discovering instances, named entities, and alternate glosses (Etzioni et al., 2005; Pasça, 2005). Additionally, a wide variety of relationship-specific classifiers have been proposed, including pattern-based classifiers for hyponyms (Hearst, 1992), meronyms (Girju, 2003), synonyms (Lin et al., 2003), a variety of verb relations (Chklovski and Pantel, 2004), and general purpose analogy relations (Turney et al., 2003). Such classifiers use hand-written or automatically-induced patterns like *Such* NP_y *as* NP_x or NP_y *like* NP_x to determine, for example that NP_y is a hyponym of NP_x (i.e., NP_y IS-A NP_x). While such classifiers have achieved some degree of success, they frequently lack the global knowledge necessary to integrate their predictions into a complex taxonomy with multiple relations.

Past work on semantic taxonomy induction includes the noun hypernym hierarchy created in (Caraballo, 2001), the part-whole taxonomies in (Girju, 2003), and a great deal of recent work described in (Buitelaar et al., 2005). Such work has typically either focused on only inferring small taxonomies over a single relation, or as in (Caraballo, 2001), has used evidence for multiple relations independently from one another, by for example first focusing strictly on inferring clusters of coordinate terms, and then by inferring hypernyms over those clusters.

Another major shortfall in previous techniques for taxonomy induction has been the inability to handle lexical ambiguity. Previous approaches have typically sidestepped the issue of polysemy altogether by making the assumption of only a single sense per word, and inferring taxonomies explicitly over words and not senses. Enforcing a false monosemy has the downside of making potentially erroneous inferences; for example, collapsing the polysemous term *Bush* into a single sense might lead one to infer by transitivity that a *rose bush* is a kind of *U.S. president*.

Our approach simultaneously provides a solution to the problems of jointly considering evidence about multiple relationships as well as lexical ambiguity within a single probabilistic framework. The key contribution of this work is to offer a solution to two crucial problems in taxonomy in-

duction and hyponym acquisition: the problem of combining heterogenous sources of evidence in a flexible way, and the problem of correctly identifying the appropriate word sense of each new word added to the taxonomy.[1]

2 A Probabilistic Framework for Taxonomy Induction

In section 2.1 we introduce our definitions for taxonomies, relations, and the *taxonomic constraints* that enforce dependencies between relations; in section 2.2 we give a probabilistic model for defining the conditional probability of a set of relational evidence given a taxonomy; in section 2.3 we formulate a local search algorithm to find the taxonomy maximizing this conditional probability; and in section 2.4 we extend our framework to deal with lexical ambiguity.

2.1 Taxonomies, Relations, and Taxonomic Constraints

We define a taxonomy $\mathbf{T}$ as a set of pairwise relations $\mathbf{R}$ over some domain of objects $\mathbf{D_T}$. For example, the relations in WordNet include *hypernymy*, *holonymy*, *verb entailment*, and many others; the objects of WordNet between which these relations hold are its word senses or *synsets*. We define that each relation $R \in \mathbf{R}$ is a set of ordered or unordered pairs of objects $(i, j) \in \mathbf{D_T}$; we define $R_{ij} \in \mathbf{T}$ if relationship R holds over objects (i, j) in $\mathbf{T}$.

Relations for Hyponym Acquisition

For the case of hyponym acquisition, the objects in our taxonomy are WordNet *synsets*. In this paper we focus on two of the many possible relationships between senses: the hypernym relation and the coordinate term relation. We treat the hypernym or ISA relation as atomic; we use the notation H^n_{ij} if a sense j is the n-th ancestor of a sense i in the hypernym hierarchy. We will simply use H_{ij} to indicate that j is an ancestor of i at some unspecified level. Two senses are typically considered to be "coordinate terms" or "taxonomic sisters" if they share an immediate parent in the hypernym hierarchy. We generalize this notion of siblinghood to state that two senses i and j are *(m, n)-cousins* if their closest *least common subsumer* (LCS)[2] is within exactly m and n links, respectively.[3] We use the notation C^{mn}_{ij} to denote that i and j are (m, n)-cousins. Thus coordinate terms are $(1, 1)$-cousins; technically the hypernym relation may also be seen as a specific case of this representation; an immediate parent in the hypernym hierarchy is a $(1, 0)$-cousin, and the k-th ancestor is a $(k, 0)$-cousin.

Taxonomic Constraints

A semantic taxonomy such as WordNet enforces certain *taxonomic constraints* which disallow particular taxonomies $\mathbf{T}$. For example, the ISA transitivity constraint in WordNet requires that each synset inherits the hypernyms of its hypernym, and the part-inheritance constraint requires that each synset inherits the meronyms of its hypernyms.

For the case of hyponym acquisition we enforce the following two taxonomic constraints on the hypernym and (m, n)-cousin relations:

1. ISA Transitivity:

$$H^m_{ij} \wedge H^n_{jk} \Rightarrow H^{m+n}_{ik}.$$

2. Definition of (m, n)-cousinhood:

$$C^{mn}_{ij} \Leftrightarrow \exists k.k = LCS(i, j) \wedge H^m_{ik} \wedge H^n_{jk}.$$

Constraint (1) requires that the each synset inherits the hypernyms of its direct hypernym; constraint (2) simply defines the (m, n)-cousin relation in terms of the atomic hypernym relation.

The addition of any new hypernym relation to a preexisting taxonomy will usually necessitate the addition of a set of other novel relations as implied by the taxonomic constraints. We refer to the full set of novel relations implied by a new link R_{ij} as $\mathbf{I}(R_{ij})$; we discuss the efficient computation of the set of implied links for the purpose of hyponym acquisition in Section 3.4.

2.2 A Probabilistic Formulation

We propose that the event $R_{ij} \in \mathbf{T}$ has some prior probability $P(R_{ij} \in \mathbf{T})$, and $P(R_{ij} \in$

[1] The taxonomies discussed in this paper are available for download at `http://ai.stanford.edu/~rion/swn`.

[2] A least common subsumer $LCS(i, j)$ is defined as a synset that is an ancestor in the hypernym hierarchy of both i and j which has no child that is also an ancestor of both i and j. When there is more than one LCS (due to multiple inheritance), we refer to the *closest* LCS, i.e.,the LCS that minimizes the maximum distance to i and j.

[3] An (m, n)-cousin for $m \geq 2$ corresponds to the English kinship relation "$(m-1)$-th cousin $|m-n|$-times removed."

$\mathbf{T}) + P(R_{ij} \notin \mathbf{T}) = 1$. We define the probability of the taxonomy as a whole as the joint probability of its component relations; given a partition of all possible relations $\mathbf{R} = \{A, B\}$ where $A \in \mathbf{T}$ and $B \notin \mathbf{T}$, we define:

$$P(\mathbf{T}) = P(A \in \mathbf{T}, B \notin \mathbf{T}).$$

We assume that we have some set of observed evidence $\mathbf{E}$ consisting of observed features over pairs of objects in some domain $\mathbf{D_E}$; we'll begin with the assumption that our features are over pairs of words, and that the objects in the taxonomy also correspond directly to words.[4] Given a set of features $E^R_{ij} \in \mathbf{E}$, we assume we have some model for inferring $P(R_{ij} \in \mathbf{T}|E^R_{ij})$, i.e., the posterior probability of the event $R_{ij} \in \mathbf{T}$ given the corresponding evidence E^R_{ij} for that relation. For example, evidence for the hypernym relation E^H_{ij} might be the set of all observed lexico-syntactic patterns containing i and j in all sentences in some corpus.

For simplicity we make the following independence assumptions: first, we assume that each item of observed evidence E^R_{ij} is independent of all other observed evidence given the taxonomy $\mathbf{T}$, i.e., $P(\mathbf{E}|\mathbf{T}) = \prod_{E^R_{ij} \in \mathbf{E}} P(E^R_{ij}|\mathbf{T})$.

Further, we assume that each item of observed evidence E^R_{ij} depends on the taxonomy $\mathbf{T}$ only by way of the corresponding relation R_{ij}, i.e.,

$$P(E^R_{ij}|\mathbf{T}) = \begin{cases} P(E^R_{ij}|R_{ij} \in \mathbf{T}) & \text{if } R_{ij} \in \mathbf{T} \\ P(E^R_{ij}|R_{ij} \notin \mathbf{T}) & \text{if } R_{ij} \notin \mathbf{T} \end{cases}$$

For example, if our evidence E^H_{ij} is a set of observed lexico-syntactic patterns indicative of hypernymy between two words i and j, we assume that whatever dependence the relations in $\mathbf{T}$ have on our observations may be explained entirely by dependence on the existence or non-existence of the single hypernym relation $H(i, j)$.

Applying these two independence assumptions we may express the conditional probability of our evidence given the taxonomy:

$$P(\mathbf{E}|\mathbf{T}) = \prod_{R_{ij} \in \mathbf{T}} P(E^R_{ij}|R_{ij} \in \mathbf{T}) \cdot \prod_{R_{ij} \notin \mathbf{T}} P(E^R_{ij}|R_{ij} \notin \mathbf{T}).$$

Rewriting the conditional probability in terms of our estimates of the posterior probabilities $P(R_{ij}|E^R_{ij})$ using Bayes Rule, we obtain:

$$P(\mathbf{E}|\mathbf{T}) = \prod_{R_{ij} \in \mathbf{T}} \frac{P(R_{ij} \in \mathbf{T}|E^R_{ij})P(E^R_{ij})}{P(R_{ij} \in \mathbf{T})} \cdot \prod_{R_{ij} \notin \mathbf{T}} \frac{P(R_{ij} \notin \mathbf{T}|E^R_{ij})P(E^R_{ij})}{P(R_{ij} \notin \mathbf{T})}.$$

Within our model we define the goal of taxonomy induction to be to find the taxonomy $\hat{\mathbf{T}}$ that maximizes the conditional probability of our observations $\mathbf{E}$ given the relationships of $\mathbf{T}$, i.e., to find

$$\hat{\mathbf{T}} = \arg\max_{\mathbf{T}} P(\mathbf{E}|\mathbf{T}).$$

2.3 Local Search Over Taxonomies

We propose a search algorithm for finding $\hat{\mathbf{T}}$ for the case of hyponym acquisition. We assume we begin with some initial (possibly empty) taxonomy $\mathbf{T}$. We restrict our consideration of possible new taxonomies to those created by the single operation ADD-RELATION($R_{ij}, \mathbf{T}$), which adds the single relation R_{ij} to $\mathbf{T}$.

We define the multiplicative change $\Delta_{\mathbf{T}}(R_{ij})$ to the conditional probability $P(\mathbf{E}|\mathbf{T})$ given the addition of a single relation R_{ij}:

$$\begin{aligned} \Delta_{\mathbf{T}}(R_{ij}) &= P(\mathbf{E}|\mathbf{T}')/P(\mathbf{E}|\mathbf{T}) \\ &= \frac{P(R_{ij} \in \mathbf{T}|E^R_{ij})P(E^R_{ij})}{P(R_{ij} \notin \mathbf{T}|E^R_{ij})P(E^R_{ij})} \cdot \frac{P(R_{ij} \notin \mathbf{T})}{P(R_{ij} \in \mathbf{T})} \\ &= k \left(\frac{P\left(R_{ij} \in \mathbf{T}|E^R_{ij}\right)}{1 - P\left(R_{ij} \in \mathbf{T}|E^R_{ij}\right)} \right). \end{aligned}$$

Here k is the inverse odds of the prior on the event $R_{ij} \in \mathbf{T}$; we consider this to be a constant independent of i, j, and the taxonomy $\mathbf{T}$.

To enforce the taxonomic constraints in $\mathbf{T}$, for each application of the ADD-RELATION operator we must add all new relations in the implied set $\mathbf{I}(R_{ij})$ not already in $\mathbf{T}$.[5] Thus we define the multiplicative change of the full set of implied relations as the product over all new relations:

$$\Delta_{\mathbf{T}}(\mathbf{I}(R_{ij})) = \prod_{R \in \mathbf{I}(R_{ij})} \Delta_{\mathbf{T}}(R).$$

[4] In section 2.4 we drop this assumption, extending our model to manage lexical ambiguity.

[5] For example, in order to add the new synset *microsoft* under the noun synset *company#n#1* in WordNet 2.1, we must necessarily add the new relations $H^2(microsoft, institution\#n\#1)$ $C^{11}(microsoft, dotcom\#n\#1)$, and so on.

This definition leads to the following best-first search algorithm for hyponym acquisition, which at each iteration defines the new taxonomy as the union of the previous taxonomy $\mathbf{T}$ and the set of novel relations implied by the relation R_{ij} that maximizes $\Delta_{\mathbf{T}}(\mathbf{I}(R_{ij}))$ and thus maximizes the conditional probability of the evidence over all possible single relations:

$$\text{WHILE} \max_{R_{ij} \notin \mathbf{T}} \Delta_{\mathbf{T}}(\mathbf{I}(R_{ij})) > 1$$
$$\mathbf{T} \leftarrow \mathbf{T} \cup \mathbf{I}(\arg\max_{R_{ij} \notin \mathbf{T}} \Delta_{\mathbf{T}}(\mathbf{I}(R_{ij}))).$$

2.4 Extending the Model to Manage Lexical Ambiguity

Since word senses are not directly observable, if the objects in the taxonomy are word senses (as in WordNet), we must extend our model to allow for a many-to-many mapping (e.g., a word-to-sense mapping) between $\mathbf{D_E}$ and $\mathbf{D_T}$. For this setting we assume we know the function $\text{senses}(i)$, mapping from the word i to all of $i's$ possible corresponding senses.

We assume that each set of word-pair evidence E^R_{ij} we possess is in fact sense-pair evidence E^R_{kl} for a specific pair of senses $k_0 \in \text{senses}(i), l_0 \in \text{senses}(j)$. Further, we assume that a new relation between two words is probable only between the correct sense pair, i.e.:

$$P(R_{kl}|E^R_{ij}) = 1\{k = k_0, l = l_0\} \cdot P(R_{ij}|E^R_{ij}).$$

When computing the conditional probability of a specific new relation $R_{kl} \in \mathbf{I}(R_{ab})$, we assume that the relevant sense pair k_0, l_0 is the one which maximizes the probability of the new relation, i.e. for $k \in \text{senses}(i), l \in \text{senses}(j)$,

$$(k_0, l_0) = \arg\max_{k,l} P(R_{kl} \in \mathbf{T}|E^R_{ij}).$$

Our independence assumptions for this extension need only to be changed slightly; we now assume that the evidence E^R_{ij} depends on the taxonomy $\mathbf{T}$ via only a single relation between sense-pairs R_{kl}. Using this revised independence assumption the derivation for best-first search over taxonomies for hyponym acquisition remains unchanged. One side effect of this revised independence assumption is that the addition of the single "sense-collapsed" relation R_{kl} in the taxonomy $\mathbf{T}$ will explain the evidence E^R_{ij} for the relation over words i and j now that such evidence has been revealed to concern only the specific senses k and l.

3 Extending WordNet

We demonstrate the ability of our model to use evidence from multiple relations to extend WordNet with novel noun hyponyms. While in principle we could use any number of relations, for simplicity we consider two primary sources of evidence: the probability of two words in WordNet being in a hypernym relation, and the probability of two words in WordNet being in a coordinate relation.

In sections 3.1 and 3.2 we describe the construction of our hypernym and coordinate classifiers, respectively; in section 3.3 we outline the efficient algorithm we use to perform local search over hyponym-extended WordNets; and in section 3.4 we give an example of the implicit structure-based word sense disambiguation performed within our framework.

3.1 Hyponym Classification

Our classifier for the hypernym relation is derived from the "hypernym-only" classifier described in (Snow et al., 2005). The features used for predicting the hypernym relationship are obtained by parsing a large corpus of newswire and encyclopedia text with MINIPAR (Lin, 1998). From the resulting dependency trees the evidence E^H_{ij} for each word pair (i, j) is constructed; the evidence takes the form of a vector of counts of occurrences that each labeled syntactic dependency path was found as the shortest path connecting i and j in some dependency tree. The labeled training set is constructed by labeling the collected feature vectors as positive "known hypernym" or negative "known non-hypernym" examples using WordNet 2.0; 49,922 feature vectors were labeled as positive training examples, and 800,828 noun pairs were labeled as negative training examples. The model for predicting $P(H_{ij}|E^H_{ij})$ is then trained using logistic regression, predicting the noun-pair hypernymy label from WordNet from the feature vector of lexico-syntactic patterns.

The hypernym classifier described above predicts the probability of the generalized hypernym-ancestor relation over *words* $P(H_{ij}|E^H_{ij})$. For the purposes of taxonomy induction, we would prefer an ancestor-distance specific set of classifiers over *senses*, i.e., for $k \in \text{senses}(i), l \in \text{senses}(j)$, the set of classifiers estimating $\{P(H^1_{kl}|E^H_{ij}), P(H^2_{kl}|E^H_{ij}), \dots\}$.

One problem that arises from directly assigning the probability $P(H_{ij}^n|E_{ij}^H) \propto P(H_{ij}|E_{ij}^H)$ for all n is the possibility of adding a novel hyponym to an overly-specific hypernym, which might still satisfy $P(H_{ij}^n|E_{ij}^H)$ for a very large n. In order to discourage unnecessary overspecification, we penalize each probability $P(H_{ij}^k|E_{ij}^H)$ by a factor λ^{k-1} for some $\lambda < 1$, and renormalize: $P(H_{ij}^k|E_{ij}^H) \propto \lambda^{k-1}P(H_{ij}|E_{ij}^H)$. In our experiments we set $\lambda = 0.95$.

3.2 (m, n)-cousin Classification

The classifier for learning coordinate terms relies on the notion of *distributional similarity*, i.e., the idea that two words with similar meanings will be used in similar contexts (Hindle, 1990). We extend this notion to suggest that words with similar meanings should be near each other in a semantic taxonomy, and in particular will likely share a hypernym as a near parent.

Our classifier for (m, n)-cousins is derived from the algorithm and corpus given in (Ravichandran et al., 2005). In that work an efficient randomized algorithm is derived for computing clusters of similar nouns. We use a set of more than 1000 distinct clusters of English nouns collected by their algorithm over 70 million webpages[6], with each noun i having a score representing its cosine similarity to the centroid c of the cluster to which it belongs, $\cos(\theta(i, c))$.

We use the cluster scores of noun pairs as input to our own algorithm for predicting the (m, n)-cousin relationship between the senses of two words i and j. If two words i and j appear in a cluster together, with cluster centroid c, we set our single coordinate input feature to be the minimum cluster score $\min(\cos(\theta(i, c)), \cos(\theta(j, c)))$, and zero otherwise. For each such noun pair feature, we construct a labeled training set of (m, n)-cousin relation labels from WordNet 2.1. We define a noun pair (i, j) to be a "known (m, n)-cousin" if for some senses $k \in \text{senses}(i), l \in \text{senses}(j)$, $C_{ij}^{mn} \in$ WordNet; if more than one such relation exists, we assume the relation with smallest sum $m + n$, breaking ties by smallest absolute difference $|m - n|$. We consider all such labeled relationships from WordNet with $0 \leq m, n \leq 7$; pairs of words that have no corresponding pair of synsets connected in the hypernym hierarchy, or with $\min(m, n) > 7$, are assigned to a single class C^∞. Further, due to the symmetry of the similarity score, we merge each class $C^{mn} = C^{mn} \cup C^{nm}$; this implies that the resulting classifier will predict, as expected given a symmetric input, $P(C_{kl}^{mn}|E_{ij}^C) = P(C_{kl}^{nm}|E_{ij}^C)$.

We find 333,473 noun synset pairs in our training set with similarity score greater than 0.15. We next apply softmax regression to learn a classifier that predicts $P(C_{ij}^{mn}|E_{ij}^C)$, predicting the WordNet class labels from the single similarity score derived from the noun pair's cluster similarity.

3.3 Details of our Implementation

Hyponym acquisition is among the simplest and most straightforward of the possible applications of our model; here we show how we efficiently implement our algorithm for this problem. First, we identify the set of all the word pairs (i, j) over which we have hypernym and/or coordinate evidence, and which might represent additions of a novel hyponym to the WordNet 2.1 taxonomy (i.e., that has a known noun hypernym and an unknown hyponym, or has a known noun coordinate term and an unknown coordinate term). This yields a list of 95,000 single links over threshold $P(R_{ij}) > 0.12$.

For each unknown hyponym i we may have several pieces of evidence; for example, for the unknown term *continental* we have 21 relevant pieces of hypernym evidence, with links to possible hypernyms {*carrier, airline, unit,* ...}; and we have 5 pieces of coordinate evidence, with links to possible coordinate terms {*airline, american eagle, airbus,* ...}.

For each proposed hypernym or coordinate link involved with the novel hyponym i, we compute the set of candidate hypernyms for i; in practice we consider all senses of the immediate hypernym j for each potential novel hypernym, and all senses of the coordinate term k and its first two hypernym ancestors for each potential coordinate.

In the *continental* example, from the 26 individual pieces of evidence over words we construct the set of 99 unique synsets that we will consider as possible hypernyms; these include the two senses of the word *airline*, the ten senses of the word *carrier*, and so forth.

Next, we iterate through each of the possible hypernym synsets l under which we might add the new word i; for each synset l we com-

[6] As a preprocessing step we hand-edit the clusters to remove those containing non-English words, terms related to adult content, and other webpage-specific clusters.

pute the change in taxonomy score resulting from adding the implied relations $\mathbf{I}(H^1_{il})$ required by the taxonomic constraints of $\mathbf{T}$. Since typically our set of all evidence involving i will be much smaller than the set of possible relations in $\mathbf{I}(H^1_{il})$, we may efficiently check whether, for each sense $s \in \text{senses}(w)$, for all words where we have some evidence E^R_{iw}, whether s participates in some relation with i in the set of implied relations $\mathbf{I}(H^1_{il})$.[7] If there is more than one sense $s \in \text{senses}(w)$, we add to $\mathbf{I}(H^1_{il})$ the single relationship R_{is} that maximizes the taxonomy likelihood, i.e. $\arg\max_{s \in \text{senses}(w)} \Delta_{\mathbf{T}}(R_{is})$.

3.4 Hypernym Sense Disambiguation

A major strength of our model is its ability to correctly choose the sense of a hypernym to which to add a novel hyponym, despite collecting evidence over untagged word pairs. In our algorithm word sense disambiguation is an implicit side-effect of our algorithm; since our algorithm chooses to add the single link which, with its implied links, yields the most likely taxonomy, and since each distinct synset in WordNet has a different immediate neighborhood of relations, our algorithm simply disambiguates each node based on its surrounding structural information.

As an example of sense disambiguation in practice, consider our example of *continental*. Suppose we are iterating through each of the 99 possible synsets under which we might add *continental* as a hyponym, and we come to the synset *airline#n#2* in WordNet 2.1, i.e. "a commercial organization serving as a common carrier." In this case we will iterate through each piece of hypernym and coordinate evidence; we find that the relation H(*continental, carrier*) is satisfied with high probability for the specific synset *carrier#n#5*, the grandparent of *airline#n#2*; thus the factor $\Delta_{\mathbf{T}}(H^3$(*continental, carrier#n#5*)) is included in the factor of the set of implied relations $\Delta_{\mathbf{T}}\left(\mathbf{I}(H^1\text{(}\textit{continental, airline\#n\#2}\text{)})\right)$.

Suppose we instead evaluate the *first* synset of *airline*, i.e., *airline#n#1*, with the gloss "a hose that carries air under pressure." For this synset none of the other 20 relationships directly implied by hypernym evidence or the 5 relationships implied by the coordinate evidence are implied by adding the single link H^1(*continental,airline#n#1*); thus the resulting change in the set of implied links given by the correct "carrier" sense of airline is much higher than that of the "hose" sense. In fact it is the largest of all the 99 considered hypernym links for *continental*; H^1(*continental, airline#n#2*) is link #18,736 added to the taxonomy by our algorithm.

[7] Checking whether or not $R_{is} \in \mathbf{I}(H^1_{il})$ may be efficiently computed by checking whether s is in the hypernym ancestors of l or if it shares a least common subsumer with l within 7 steps.

4 Evaluation

In order to evaluate our framework for taxonomy induction, we have applied hyponym acquisition to construct several distinct taxonomies, starting with the base of WordNet 2.1 and only adding novel noun hyponyms. Further, we have constructed taxonomies using a baseline algorithm, which uses the identical hypernym and coordinate classifiers used in our joint algorithm, but which does not combine the evidence of the classifiers.

In section 4.1 we describe our evaluation methodology; in sections 4.2 and 4.3 we analyze the fine-grained precision and disambiguation precision of our algorithm compared to the baseline; in section 4.4 we compare the coarse-grained precision of our links (motivated by categories defined by the WordNet *supersenses*) against the baseline algorithm and against an "oracle" for named entity recognition.

Finally, in section 4.5 we evaluate the taxonomies inferred by our algorithm directly against the WordNet 2.1 taxonomy; we perform this evaluation by testing each taxonomy on a set of human judgments of hypernym and non-hypernym noun pairs sampled from newswire text.

4.1 Methodology

We evaluate the quality of our acquired hyponyms by direct judgment. In four separate annotation sessions, two judges labeled $\{$50,100,100,100$\}$ samples uniformly generated from the first $\{$100,1000,10000,20000$\}$ single links added by our algorithm.

For the direct measure of fine-grained precision, we simply ask for each link $H(X, Y)$ added by the system, is X a Y? In addition to the fine-grained precision, we give a coarse-grained evaluation, inspired by the idea of supersense-tagging in (Ciaramita and Johnson, 2003). The 26 supersenses used in WordNet 2.1 are listed in Table 1; we label a hyponym link as correct in the coarse-grained evaluation if the novel hyponym is placed under the appropriate supersense. This evaluation task

1 Tops	8 communication	15 object	22 relation
2 act	9 event	16 person	23 shape
3 animal	10 feeling	17 phenomenon	24 state
4 artifact	11 food	18 plant	25 substance
5 attribute	12 group	19 possession	26 time
6 body	13 location	20 process	
7 cognition	14 motive	21 quantity	

Table 1: The 26 WordNet supersenses

is similar to a fine-grained Named Entity Recognition (Fleischman and Hovy, 2002) task with 26 categories; for example, if our algorithm mistakenly inserts a novel non-capital city under the hyponym *state capital*, it will inherit the correct supersense *location*. Finally, we evaluate the ability of our algorithm to correctly choose the appropriate sense of the hypernym under which a novel hyponym is being added. Our labelers categorize each candidate sense-disambiguated hypernym synset suggested by our algorithm into the following categories:

c_1**:** Correct sense-disambiguated hypernym.

c_2**:** Correct hypernym word, but incorrect sense of that word.

c_3**:** Incorrect hypernym, but correct supersense.

c_4**:** Any other relation is considered incorrect.

A single hyponym/hypernym pair is allowed to be simultaneously labeled 2 and 3.

4.2 Fine-grained evaluation

Table 2 displays the results of our evaluation of fine-grained precision for the baseline non-joint algorithm (**Base**) and our joint algorithm (**Joint**), as well as the relative error reduction (**ER**) of our algorithm over the baseline. We use the minimum of the two judges' scores. Here we define fine-grained precision as $c_1/total$. We see that our joint algorithm strongly outperforms the baseline, and has high precision for predicting novel hyponyms up to 10,000 links.

4.3 Hypernym sense disambiguation

Also in Table 2 we compare the sense disambiguation precision of our algorithm and the baseline. Here we measure the precision of sense-disambiguation among all examples where each algorithm found a correct hyponym word; our calculation for disambiguation precision is $c_1/(c_1 + c_2)$. Again our joint algorithm outperforms the baseline algorithm at all levels of recall. Interestingly the baseline disambiguation precision improves with higher recall; this may be attributed to the observation that the highest-confidence hypernyms predicted by individual classifiers are likely to be polysemous, whereas hypernyms of lower confidence are more frequently monosemous (and thus trivially easy to disambiguate).

	Fine-grained Pre.			**Disambiguation Pre.**		
#Links	Base	Joint	ER	Base	Joint	ER
100	0.60	1.00	100%	0.86	1.00	100%
1000	0.52	0.93	85%	0.84	1.00	100%
10000	0.46	0.84	70%	0.90	1.00	100%
20000	0.46	0.68	41%	0.94	0.98	68%

Table 2: Fine-grained and disambiguation precision and error reduction for hyponym acquisition

# Links	NER Oracle	Base	Joint	ER vs. NER	ER vs. Base
100	1.00	0.72	1.00	0%	100%
1000	0.69	0.68	0.99	97%	85%
10000	0.45	0.69	0.96	93%	70%
20000	0.54	0.69	0.92	83%	41%

Table 3: Coarse-grained precision and error reduction vs. Non-joint baseline and NER Oracle

4.4 Coarse-grained evaluation

We compute coarse-grained precision as $(c_1 + c_3)/total$. Inferring the correct coarse-grained supersense of a novel hyponym can be viewed as a fine-grained (26-category) Named Entity Recognition task; our algorithm for taxonomy induction can thus be viewed as performing high-accuracy fine-grained NER. Here we compare against both the baseline non-joint algorithm as well as an "oracle" algorithm for Named Entity Recognition, which perfectly classifies the supersense of all nouns that fall under the four supersenses $\{person, group, location, quantity\}$, but works only for those supersenses. Table 3 shows the results of this coarse-grained evaluation. We see that the baseline non-joint algorithm has higher precision than the NER oracle as 10,000 and 20,000 links; however, both are significantly outperformed by our joint algorithm, which maintains high coarse-grained precision (92%) even at 20,000 links.

4.5 Comparison of inferred taxonomies and WordNet

For our final evaluation we compare our learned taxonomies directly against the currently existing hypernym links in WordNet 2.1. In order to compare taxonomies we use a hand-labeled test

	WN	+10K	+20K	+30K	+40K
PRE	0.524	0.524	0.574	**0.583**	0.571
REC	0.165	0.165	0.203	**0.211**	0.211
F	0.251	0.251	0.300	**0.309**	0.307

Table 4: Taxonomy hypernym classification vs. WordNet 2.1 on hand-labeled testset

set of over 5,000 noun pairs, randomly-sampled from newswire corpora (described in (Snow et al., 2005)). We measured the performance of both our inferred taxonomies and WordNet against this test set.[8] The performance and comparison of the best WordNet classifier vs. our taxonomies is given in Table 4. Our best-performing inferred taxonomy on this test set is achieved after adding 30,000 novel hyponyms, achieving an 23% relative improvement in F-score over the WN2.1 classifier.

5 Conclusions

We have presented an algorithm for inducing semantic taxonomies which attempts to globally optimize the entire structure of the taxonomy. Our probabilistic architecture also includes a new model for learning coordinate terms based on (m, n)-cousin classification. The model's ability to integrate heterogeneous evidence from different classifiers offers a solution to the key problem of choosing the correct word sense to which to attach a new hypernym.

Acknowledgements

Thanks to Christiane Fellbaum, Rajat Raina, Bill MacCartney, and Allison Buckley for useful discussions and assistance annotating data. Rion Snow is supported by an NDSEG Fellowship sponsored by the DOD and AFOSR. This work was supported in part by the Disruptive Technology Office (DTO)'s Advanced Question Answering for Intelligence (AQUAINT) Program.

References

P. Buitelaar, P. Cimiano and B. Magnini. 2005. Ontology Learning from Text: Methods, Evaluation and Applications. Volume 123 Frontiers in Artificial Intelligence and Applications.

S. Caraballo. 2001. Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text. Brown University Ph.D. Thesis.

S. Cederberg and D. Widdows. 2003. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. *Proc. CoNLL-2003*, pp. 111–118.

T. Chklovski and P. Pantel. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. *Proc. EMNLP-2004*.

M. Ciaramita and M. Johnson. 2003. Supersense Tagging of Unknown Nouns in WordNet. *Proc. EMNLP-2003*.

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.

C. Fellbaum. 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

R. Girju, A. Badulescu, and D. Moldovan. 2003. Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. *Proc. HLT-03*.

M. Fleischman and E. Hovy. 2002. Fine grained classification of named entities. *Proc. COLING-02*.

M. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. *Proc. COLING-92*.

D. Hindle. 1990. Noun classification from predicate-argument structures. *Proc. ACL-90*.

D. Lenat. 1995. CYC: A Large-Scale Investment in Knowledge Infrastructure, Communications of the ACM, 38:11, 33–35.

D. Lin. 1998. Dependency-based Evaluation of MINIPAR. *Workshop on the Evaluation of Parsing Systems, Granada, Spain*.

D. Lin, S. Zhao, L. Qin and M. Zhou. 2003. Identifying Synonyms among Distributionally Similar Words. Proc. IJCAI-03.

M. Pasça. 2005. Finding Instance Names and Alternative Glosses on the Web: WordNet Reloaded. *CICLing 2005*, pp. 280-292.

D. Ravichandran, P. Pantel, and E. Hovy. 2002. Randomized Algorithms and NLP: Using Locality Sensitive Hash Function for High Speed Noun Clustering. *Proc. ACL-2002*.

E. Riloff and J. Shepherd. 1997. A Corpus-Based Approach for Building Semantic Lexicons. *Proc EMNLP-1997*.

B. Roark and E. Charniak. 1998. Noun-phrase co-occurerence statistics for semi-automatic-semantic lexicon construction. *Proc. ACL-1998*.

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

P. Turney, M. Littman, J. Bigham, and V. Shnayder. 2003. Combining independent modules to solve multiple-choice synonym and analogy problems. *Proc. RANLP-2003*, pp. 482–489.

[8] We found that the WordNet 2.1 model achieving the highest F-score used only the first sense of each hyponym, and allowed a maximum distance of 4 edges between each hyponym and its hypernym.