# Automatically Inducing Ontologies from Corpora

**Inderjeet Mani**

**Department of Linguistics**
**Georgetown University, ICC 452**
**37th and O Sts, NW**
**Washington, DC 20057, USA**
im5@georgetown.edu

**Ken Samuel, Kris Concepcion and David Vogel**

**The MITRE Corporation**
**7515 Colshire Drive**
**McLean, VA 22102, USA**
{samuel, kjc9, dvogel}@mitre.org

**Abstract**

The emergence of vast quantities of on-line information has raised the importance of methods for automatic cataloguing of information in a variety of domains, including electronic commerce and bioinformatics. Ontologies can play a critical role in such cataloguing. In this paper, we describe a system that automatically induces an ontology from any large on-line text collection in a specific domain. The ontology that is induced consists of domain concepts, related by *kind-of* and *part-of* links. To achieve domain-independence, we use a combination of relatively shallow methods along with any available repositories of applicable background knowledge. We describe our evaluation experiences using these methods, and provide examples of induced structures.

## 1 Introduction

The emergence of vast quantities of on-line information has raised the importance of methods for automatic cataloguing of information in a variety of domains, including electronic commerce and bioinformatics. Ontologies[1] can play a critical role in such cataloguing. In bioinformatics, for example, there is growing recognition that common ontologies, e.g., the Gene Ontology[2], are critical to interoperation and integration of biological data, including both structured data as found in protein databases, as well as unstructured data, as found in on-line biomedical literature.

Constructing an ontology is an extremely laborious effort. Even with some reuse of "core" knowledge from an Upper Model (Cohen et al. 1999), the task of creating an ontology for a particular domain and task has a high cost, incurred for each new domain. Tools that could automate, or semi-automate, the construction of ontologies for different domains could dramatically reduce the knowledge creation cost.

One approach to developing such tools is to rely on information implicit in collections of on-line text in a particular domain. If it were possible to automatically extract terms and their semantic relations from the text corpus, the ontology developer could build on that knowledge, revising it, as needed, etc. This would be more cost-effective than having a human develop the ontology from scratch.

Our approach is inspired by research on topic-focused multi-document summarization of large text collections, where there is a need to characterize the collection content succinctly in a hierarchy of topic terms and their relationships. Current approaches to multi-document summarization combine linguistic analysis, corpus statistics, and the use of background semantic knowledge from generic thesauri such as WordNet to infer semantic information about a person. In extending such approaches to ontology induction, the hypothesis is that similar hybrid approaches can be used to identify technical terms in a domain-specific corpus and infer semantic relationships among them.

In this paper, we describe a system that automatically induces an ontology from any large on-line text collection in a specific domain, to support cataloguing in information access and data integration tasks. The induced ontology consists of domain concepts related by *kind-of* and *part-of* links, but does not include more specialized relations or axioms. The structure of the ontology is a directed acyclic graph (DAG). To achieve domain-independence, we use a combination of relatively shallow methods along with existing repositories of applicable background knowledge. These are described in Section 2. In Section 3, we also introduce a new metric *Relation Precision* for evaluating induced ontologies in comparison with reference ontologies. We have applied our system to produce ontologies in numerous domains:

[1] This research was supported by the National Science Foundation (ITR-0205470).

[2] www.geneontology.org

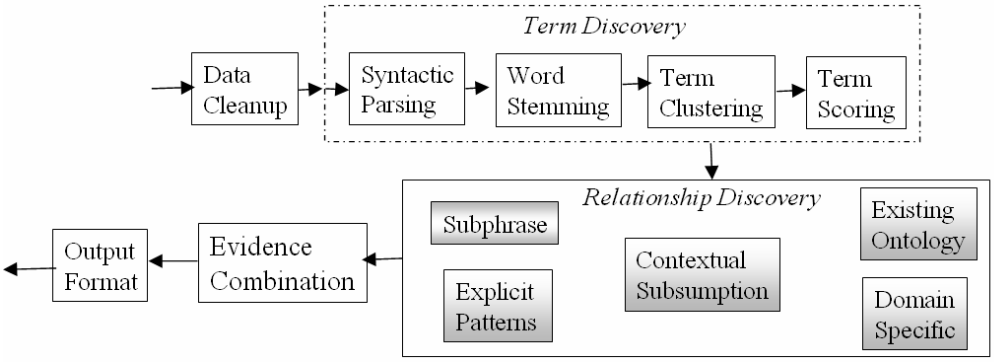


**Figure 1: System Architecture**

| | | | |
|---|---|---|---|
| **IRS Publication 17** | 285<br>$k_1$ | 0 | 285<br>$n_1$ |
| **Reuters Corpus** | 9<br>$k_2$ | 19,024 | 19,043<br>$n_2$ |
| **Total** | 294 | 19,024 | 19,328 |

**Table 1: Distribution of 'income tax' in domain and background corpora**

(i) newswire from the TREC collection (ii) taxation information from the IRS (Publication 17, from (IRS 2001)), (iii) epidemiological newsgroup messages from the Program for Monitoring Emerging Diseases (PROMED) from the Federation of American Scientists[3], (iv) the text of a book by the first author called *Automatic Summarization*, and (v) MEDLINE biomedical abstracts retrieved from the National Library of Medicine's PubMed system[4]. In the latter domain, we have begun building a large ontology using the ontology induction methods along with post-editing by domain experts in molecular biology at Georgetown University[5]. This ontology, called PRONTO, involves hundreds of thousands of protein names found in MEDLINE abstracts and in UNIPROT, the world's largest protein database[6]. It is therefore infeasible to construct PRONTO by hand from scratch. PRONTO is also much larger than other ontologies in the biology area; for example, the Gene Ontology is rather high-level, and contains (as of March 2004) only about 17,000 terms.

[3] www.fas.org/promed/

[4] www4.ncbi.nlm.nih.gov/PubMed/

[5] complingone.georgetown.edu/~prot/

[6] pir.georgetown.edu

## 2 Approach

### 2.1 System Architecture

An overall architecture for domain-independent ontology induction is shown in Figure 1. The documents are preprocessed to separate out headers. Next, terms are extracted using finite-state syntactic parsing and scored to discover domain-relevant terms. The subsequent processing infers semantic relations between pairs of terms using the 'weak' knowledge sources run in the order described below. Evidence from multiple knowledge sources is then combined to infer the resulting relations. The resulting ontologies are written out in a standard XML-based format (e.g., XOL, RDF, OWL), for use in various information access applications.

While the ontology induction procedure does not involve human labor, except for writing the preprocessing and term tokenization program for specialized technical domains, the human may edit the resulting ontology for use in a given application. An ontology editor has been developed, discussed briefly in Section 3.1.

### 2.2 Term Discovery

The system takes a collection of documents in a subject area, and identifies terms characteristic of the domain. In a given domain such as

bioinformatics, specialized term tokenization (into single- and multi-word terms) is required. The protein names can be long, e.g., "steroid/thyroid/retinoic nuclear hormone receptor homolog nhr-35", and involve specialized patterns. In constructing PRONTO, we have used a protein name tagger based on an ensemble of statistical classifiers to tag protein names in collections of MEDLINE abstracts (Anon 2004). Thus, in such a domain, a specialized tagger replaces the components in the dotted box in Figure 1.

In other domains, we adopt a generic term-discovery approach. Here the text is tagged for part-of-speech, and single- and multi-word terms consisting of minimal NPs are extracted using finite-state parsing with CASS (Abney 1996). All punctuation except for hyphens are removed from the terms, which are then lower-cased. Each word in each term is stemmed, with statistics (see below) being gathered for each stemmed term. Multi-word terms are clustered so that open, closed and hyphenated compounds are treated as equivalent, with the most frequent term in the collection being used as the cluster representative.

The terms are scored for domain-relevance based on the assumption that if a term occurs significantly more in a domain corpus than in a more diffuse background corpus, then the term is clearly domain relevant.

As an illustration, in Table 1 we compare the number of documents containing the term 'income tax' (or 'income taxes') in a long (2.18 Mb) IRS publication, Publication 17, from an IRS web site (IRS 2001) compared to a larger (27.63 Mb subset of the) Reuters 21578 news corpus[7]. One would expect that 'income tax' is much more a characteristic of the IRS publication, and this is borne out by the document frequencies in the table. We use the log likelihood ratio (LLR) (Dunning 1993) given by

$$-2\log_2(H_o(p;k_1,n_1,k_2,n_2)/H_a(p_1,p_2;n_1,k_1,n_2,k_2))$$

LLR measures the extent to which a hypothesized model of the distribution of cell counts, $H_a$, differs from the null hypothesis, $H_o$ (namely, that the percentage of documents containing this term is the same in both corpora). We used a binomial model for $H_o$ and $H_a$[8].

### 2.3 Relationship Discovery

The main innovation in our approach is to fuse together information from multiple knowledge sources as evidence for particular semantic relationships between terms. To infer semantic relations such as *kind-of* and *part-of*, the system uses a bottom-up data-driven approach using a combination of evidence from shallow methods.

#### 2.3.1 Subphrase Relations

These are based on the presence of common syntactic heads, and allow us to infer, for example, that 'p68 protein' is a *kind-of* 'protein'. Likewise, in the TREC domain, subphrase analysis tells us that 'electric car' is a kind of 'car', and in the IRS domain, that 'federal income tax' is a kind of 'income tax'.

#### 2.3.2 Existing Ontology Relations

These are obtained from a thesaurus. For example, the Gene Ontology can be used to infer that 'ATP-dependent RNA helicase' is a *kind of* 'RNA-helicase'. Likewise, in the TREC domain, using WordNet tells us that 'tailpipe' is *part of* 'automobile', and in the IRS domain, that 'spouse' is a *kind of* 'person'. Synonyms are also merged together at this stage.

#### 2.3.3 Contextual Subsumption Relations

We also infer hierarchical relations between terms, by top-down clustering using a context-based subsumption (CBS) algorithm. The algorithm uses a probabilistic measure of set covering to find subsumption relations. For each term in the corpus, we note the set of contexts in which the term appears. Term1 is said to subsume term2 when the conditional probability of term1 appearing in a context given the presence of term2, i.e., P(term1|term2), is greater than some threshold.

CBS is based on the algorithm of (Lawrie et al. 2001), which used a greedy approximation of the Domination Set Problem for graphs to discover subsumption relations among terms. Unlike their work, we did not seek to minimize the set of covering terms; therefore, a subsumed term may have multiple parents. The conditional probability threshold (0.8) we use to determine subsumption is much higher than in their approach. We also restrict the height of the hierarchies we build to three tiers. Tightening these latter two constraints appears to notably improve the quality of our subsumption relations.

The largest corpus against which CBS has run is the ProMed corpus where, considering each paragraph a distinct context, there were 117,690 contexts in the 11,198 documents. Here is an example from ProMed of a transitive relation that spans three tiers: 'mosquito' is a hypernym of 'mosquito pool', and 'mosquito' is also a hypernym of 'standing water'.

[7] In Publication 17, each "chapter" is a document.

[8] From Table 1, $p=294/19238=.015$, $p_1=285/285=1.0$, $p_2=9/19043=4.72$, $k_1=285$, $n_1=285$, $k_2=9$, $n_2=19043$.

### 2.3.4 Explicit Patterns Relations

This knowledge source infers specific relations between terms based on characteristic *cue-phrases* which relate them. For example, the cue-phrase "such as" (Hearst 1992) (Caraballo 1999) suggest a *kind-of* relation, e.g., 'a ligand such as triethylphosphine' tells us that 'triethylphosphene' is a kind of 'ligand'. Likewise, in the TREC domain, 'air toxics such as benzene' can suggest that 'benzene' is a *kind of* 'air toxic'. However, since such cue-phrase patterns tend to be sparse in occurrence, we do not use them in the evaluations described below.

### 2.3.5 Domain-Specific Knowledge Sources

Although our approach is domain-independent, it is possible to factor in domain knowledge sources for a given domain. For example, in biology, 'ase' is usually a suffix indicating an enzyme. Postmodifying PPs (found using a CASS grammar) can also be useful in some domains, as shown in 'tax on investment income of child' in Figure 2. We have so far, however, not investigated other domain-specific knowledge sources.

## 2.4 Evidence Combination

The main point about these and other knowledge sources is that each may provide only partial information. Combining these knowledge sources together, we expect, will lead to superior performance compared to just any one of them. Not only do inferences from different knowledge sources support each other, but they are also combined to produce new inferences by transitivity relations. For example, since phrase analysis tells us that 'pyridine metabolism' is a *kind-of* 'metabolism', and Gene Ontology tells us that 'metabolism' is a *kind-of* 'biological process', it follows that 'pyridine metabolism' is a *kind-of* 'biological process'. The evidence combination, in addition to computing transitive closure of these relations, also detects inconsistencies, querying the user to resolve them when detected.

# 3 Evaluation

## 3.1 Informal Assessment

Subphrase Relations is a relatively high-precision knowledge source compared to the others, producing many linked chains. Its performance can be improved by flagging and excluding proper names and idioms from its input (e..g, so that 'palm pilot' doesn't show up as a kind-of 'pilot'). However, a chain of such relations can be interrupted by terms that aren't lexically similar, but that are nevertheless in a kind-of relation. Some of these gaps are filled by transitivity relations involving other knowledge sources, especially Existing Ontologies, which is especially useful in filling gaps in some of the upper levels of the ontology. While Contextual Subsumption is good at discovering associations between 'leaves' in the DAG and other concepts, the method cannot reliably infer the label of the relation. For example, in the IRS domain, we obtain 'divorce' as more general than 'decree of divorce' and 'separate maintenance', but we don't know the nature of the relations. Contextual Subsumption-inferred links are directed edges with label 'unknown'.

Overall, the ontologies produced are noisy and require human correction, and the methods can produce many fragments that need to be linked by hand. While the system can detect cycles that need resolution by the human, these rarely arise

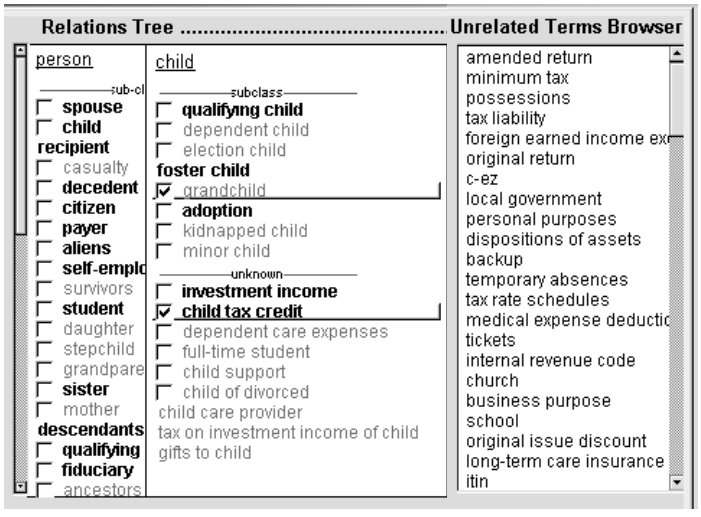


**Figure 2: An IRS Ontology viewed in the Ontology Editor**

| Term | Target DF | Back-ground DF | LLR | IG | MI | DF | TF | TF * IDF |
|---|---|---|---|---|---|---|---|---|
| **electric** | 80 | 61 | 99.9 | 99.9 | 81.3 | 99.9 | 99.9 | 27.8 |
| **car** | 77 | 56 | 99.6 | 99.3 | 81.5 | 99.8 | 99.9 | 79.4 |
| **battery** | 54 | 16 | 99.0 | 98.2 | 86.9 | 98.7 | 99.9 | 94.9 |
| **emission** | 15 | 0 | 96.5 | 96.8 | 99.2 | 79.1 | 96.6 | 64.8 |
| year | 58 | 505 | 67.9 | 67.6 | 25.0 | 99.2 | 99.7 | 65.7 |
| informal | 10 | 29 | 66.2 | 66.3 | 0.2 | 48.6 | 99.7 | 99.2 |
| record | 8 | 138 | 15.2 | 15.7 | 4.4 | 50.2 | 99.9 | 99.9 |
| osha | 1 | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 99.9 | 0.0 |

**Table 2: Comparing Topic 230 Term Percentile Rankings**

For a flavor of the kind of results we get, see Figure 2, which displays an ontology induced without any human intervention from IRS Publication 17. Here the DAG is displayed as a tree. The immediate children of 'person', a node high in the ontology, is shown in the left part of the window. Selecting 'child' brings up its kinds as well as some other children linked by "unknown" label via Contextual Subsumption, e.g., 'full-time student'. A list of orphaned terms that aren't related to any others are shown on the far right. The terms with checkboxes are those that occur in the corpus; the others are those that are found exclusively by Existing Ontology Relations. Checking a term allows it to be inspected in its occurrence context in the corpus. The editor comes with a variety of tools to help integrate ontology fragments.

### 3.2 Human Evaluation

#### 3.2.1 Term Scoring

To evaluate term scoring, we used a corpus of news articles about automobiles that consisted of 85 documents relevant to the TREC Topic 230 query: "Is the automobile industry making an honest effort to develop and produce an electric-powered automobile?" In Table 2, we provide some examples of how the LLR term scoring statistic performed with respect to five others on selected unigrams in the Topic 230 domain: term frequency, document frequency, term frequency times inverse document frequency (TF*IDF), pointwise mutual information (MI), and information gain (IG). Terms in bold are ones we judged important in the Topic 230 domain, the others are deemed unimportant. The numbers are percentile rankings. LLR and IG do equally well, outperforming the others.

We carried out other comparisons for two other domains. In the income-tax domain, a hand-built term list from the IRS contained 82 terms which occurred in IRS Publication 17, of which the system discovered 77 (94% recall). In the ProMed domain, a pre-existing hand-built taxonomy produced by a bioterrorism analyst had 1048 terms which occurred in the ProMed message corpus, of which 607 were discovered by the system (58% recall). However, the hand-built taxonomy, which was built without consulting a corpus, wasn't a full-fledged ontology, for example, there was no label for the parent-child relation.

#### 3.2.2 Term Relationships

We also carried out an evaluation experiment to determine if the relations being discovered by the machine were in keeping with human judgments. We focused here on an evaluation of pairs of knowledge sources. Our experiment examined the case where the system discovered a *kind-of* relation. Here each subject was first asked to read four newspaper articles from the TREC topic-230 sub-collection. The articles were then kept accessible to the subject in a browser window for the subject to consult if needed in answering subsequent questions. The subject was asked to judge, based on the documents read, whether term X was a kind of term Y, term Y was a kind of term X, or neither; e.g., "Is *acid* a kind of *pollutant*, or is *pollutant* a kind of *acid*, or neither?". The subject had one of three mutually exclusive choices; the first two choices were presented in randomized order.

The subjects were 16 native speakers of English unconnected with the project. Each subject was given ten questions to answer in each of the experiments. For each set of ten questions, five were chosen at random from pairs of terms related

by (immediate) *kind-of* relations. The remaining five questions were chosen at random from pairs of terms between which the system found no relation whatsoever.

| | Human | |
|---|---|---|
| **System** | *kind-of(A, B)* | *not kind-of(A,B)* |
| *kind-of(A, B)* | 56 | 18 |
| *not kind-of(A,B)* | 6 | 74 |

**Table 3: Is X a kind-of Y?**

We first discuss inter-subject agreement. Three subjects given the same relation to judge agreed 75% of the time, leading to a Kappa score of 0.72, indicating a good level of agreement. This means that subjects were able to reliably make judgments as to whether A is a *kind of* B in some document.

The results for the 16 subjects are shown in Table 3. When the system is compared to the human as ground truth, this gives a Precision of .90, a Recall of .75, and an F-measure of .82. This performance is also significantly better than random assignment: with chi-square=74.29, with $p < 0.001$[9]. The substantial effect sizes of the chi-square indicates a very solid result. There were 62 decisions involving Subphrase Relations (with 44 True Positives and 18 False Negatives), and 10 decisions involving WordNet (with 12 True Positives). This shows that there is solid agreement between the human subjects and the system on the *kind-of* relations. However, these 154 decisions involved only four newspaper articles, so clearly more data would be helpful.

### 3.3 Automatic Evaluation

While evaluation by humans is valuable, it is expensive to carry out, and this expense must be incurred each time one wants to do an evaluation. Automatic comparison of a machine-generated ontology against reference ontologies constructed by humans, e.g., (Zhang et al. 1996) (Sekine et al. 1999) (Daude et al. 2001), is therefore desirable, provided suitable reference ontologies are available. In this evaluation, the human-generated taxonomy for ProMed described in Section 3.2.1 was used as a reference ontology, with its unlabeled parent-child relation treated as a *kind-of* link. However, the human 'ontology' was created without looking at a corpus, and was developed for use with a different set of goals in mind. Although this involves comparing 'apples' and 'oranges', a comparison is nevertheless illustrative, and can in addition be useful when comparing mutiple ontologies created under similar conditions.

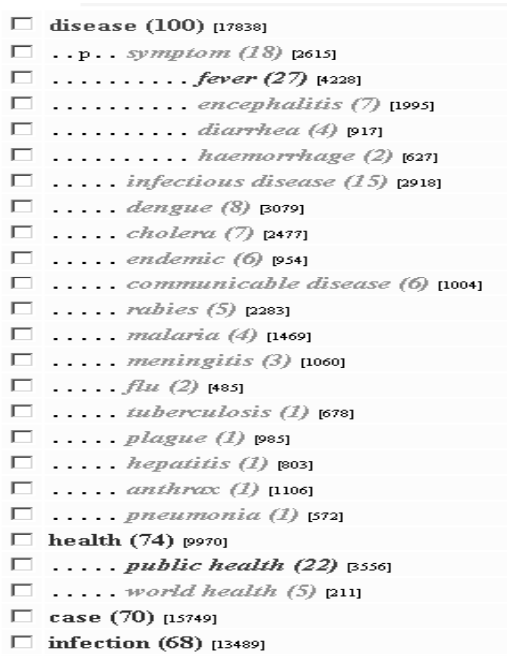


**Figure 3: Automatically Induced Fragment from ProMed**

To set aside the problem of differences in terminology involved in the comparison, we decided to restrict our attention to the set of terms $T_H$ (of cardinality 3025) in the human ontology (H), and have our system induce relations between them using the ProMed corpus. Relations were induced automatically in the machine ontology (M) for just 761 of those terms, yielding a set $T_{H1}$. The structure of $T_{H1}$ is shown in a fragment in Figure 3. Here A is a *kind-of* B if it is printed under B without a label; A is a *part-of* B if it is printed under B with a "p" label.

We then automatically computed, for each pair of terms $t_1$ and $t_2$ in $T_{H1}$ that were linked distance 1 apart in M, the distance between those terms in H. Likewise, we also computed, for each pair of terms $t_1$ and $t_2$ in $T_{H1}$ distance 1 apart in H, the distance between those terms in M.

The results of this comparison are as follows. The number of relations where the two ontologies agree exactly is 63 (i.e., the terms are distance 1 apart in both ontologies). Since, given a set of terms, there are many different ways to construct an ontology, this is encouraging.

The number of relations that our system found which were 'missed', i.e., more than distance 1 away, in H is 1203. Given the previous experiment where the human subjects agreed with the system's relations, these 1203 relations are likely to contain many that the human probably missed. For example, the relations in the machine ontology between 'eye' and 'farsightedness', and 'medicine'

[9] The chi-square for Subphrase Relations is 61.68, and the chi-square for WordNet is 56.73, with $p < 0.001$ in all cases.

and 'chiropractic medicine' are missed by H. This highlights a problem with human-generated ontologies: substantial errors of omission.

The number of relations in H that our system missed (relations that were more than distance 1 away in the system ontology), is 3493. However, of these 3493 relations, 2955 involved at least 1 term that was not included in M, leaving 538 relations that we could calculate the distance for in M. These 538 relations in H include relations between 'acid indigestion medicine' and 'maalox', and 'alternative medicine' and 'acupuncture' (a majority of the misses involved relations between a disease and the name of a specific drug for it, which aren't *part-of* or *kind-of* relations).

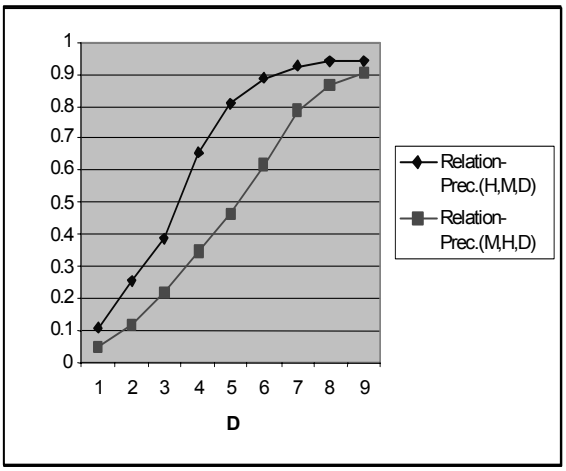


**Figure 4: Relation Precision**

These observations lead to a metric for comparing one ontology with another one serving as a reference ontology. Given two ontologies A and B, define *Relation Precision (A, B, D)* as t*he proportion of the distance 1 relations in A that are at most a distance D apart in B*. This measure can be plotted for different values of D. In Figure 4, we show the Relation Precision(H, M, D), and Relation Precision(M, H, D), for our machine ontology M and human ontology H. Both curves show Relation Precision(H, M, D) growing faster than Relation Precision(M, H, D), with 70% of the area being below the former curve and 54% being below the latter curve. The graph shows that while 22% of distance 1 relations in M are at most 3 apart in H (but keep in mind the errors of omission in H), 40% of distance 1 relations in H are at most 3 apart in M[10].

[10] The mean distance in H between terms that are distance 1 apart in M is 5.17, with a standard deviation of 2.12. The mean distance in M between terms which are distance 1 apart in H is 3.85, with a standard deviation of 1.69.

## 4 Related Work

The existing approaches to ontology induction include those that start from structured data, merging ontologies or database schemas (Doan et al. 2002). Other approaches use natural language data, sometimes just by analyzing the corpus (Sanderson and Croft 1999), (Caraballo 1999) or by learning to expand WordNet with clusters of terms from a corpus, e.g., (Girju et al. 2003). Information extraction approaches that infer labeled relations either require substantial hand-created linguistic or domain knowledge, e.g., (Craven and Kumlien 1999) (Hull and Gomez 1993), or require human-annotated training data with relation information for each domain (Craven et al. 1998).

Many, though not all, domain-independent approaches (Evans et al. 1991) (Grefenstette 1997) have restricted themselves to discovering term-associations, rather than labeled relations. A notable exception is (Sanderson and Croft 1995), which (unlike our approach) assumes the existence of a query that was used to originally retrieve the documents (so that terms can be extracted from the query and then expanded to generate additional terms for the ontology). Their approach also is restricted to one method to discover relations, while we use several.

Our approach is complementary to approaches aimed at automatically enhancing existing resources for a particular domain, e.g. (Moldovan et al. 2000). Finally, the prior methods, while they often carry out evaluation, lack standard criteria for ontology evaluation. Although ontology evaluation remains challenging, we have discussed several evaluation methods in this paper.

## 5 Conclusion

The evidence combination described above is based on transitivity and union. Since the above evaluations, we have been experimenting with an ad hoc weighted evidence combination scheme, based on each knowledge source expressing a strength for a posited relation. In future, we will also investigate using an initial seed ontology to provide a better 'backbone' for induction, and then using a spreading activation method to activate nodes related by existing knowledge sources to seed nodes. Corpus statistics can be used to weight the links. For example, based on (Caraballo 1999), each parent of a leaf node could be viewed as a cluster label for its children, with the weight of a parent-child link being determined based on how strongly the child is associated with the cluster.

The ontology induction methods described here can allow for considerable savings in time in

constructing ontologies. The evaluations we have carried out are suggestive, but many issues remain open. There are many unanswered questions about human-created reference ontologies, including lack of inter-annotator agreement studies. Indeed, experience shows that without guidelines for ontology construction, humans are prone to come up with very different ontologies for a domain. Comparing a machine-induced ontology against an ideal human reference ontology, were one to be available, is also fraught with problems. Our experience with using an implementation of the (Daude et al. 2001) constraint relaxation algorithm for ontology comparison suggests that much work is needed on distance metrics which are not over-sensitive to small differences in structure.

Our interest, therefore, is focused more towards an extrinsic evaluation. PRONTO, which is due to be released in 2004, offers the opportunity to measure costs of ontology induction and post-editing on a large-scale problem of value to the biology community. We also plan to measure the effectiveness of PRONTO in query expansion for information access to MEDLINE and protein databases. Finally, we will investigate more sophisticated evidence combination methods, and compare against other automatic methods for ontology induction.

The ontology induction tools are available for free distribution for research purposes.

**References**


Abney, S. 1996. Partial parsing Via Finite-State Cascades. *Proceedings of the ESSLLI '96 Robust Parsing Workshop*.

Caraballo, S. A. 1999. Automatic Construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (ACL'1999), 120-122.

Cohen, P. R., Chaudhri, V., Pease, A. and Schrag, R. 1999. Does Prior Knowledge Facilitate the Development of Knowledge-based Systems? The Sixteenth National Conference on Artificial Intelligence (AAAI-99).

Craven, M. and Kumlien, J. 1999. Constructing biological knowledge bases by extracting information from text sources. *Proc Int Conf Intell Syst Mol Biol.*, 77-86.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., and Slattery, S.. 1998. Learning to Extract Symbolic Knowledge from the World Wide Web. Proceedings of AAAI-98, 509-516.

Daude, J., Padro, L. and Rigau, G. 2001 A Complete WN1.5 to WN1.6 Mapping. *NAACL-2001 Workshop on WordNet and Other Lexical Resources: Applications, Extension, and Customization*, 83-88.

Doan, A., Madhavan, J. , Domings, P. and Halevy, A. 2002. Learning to Map between Ontologies on the Semantic Web. WWW'2002.

Dunning, T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, 19(1):61-74, 1993.

Girju, R., Badulescu, A., and Moldovan, D. 2003. Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. Proceedings of HLT'2003, Edmonton.

Grefenstette, G. 1997. *Explorations in Automatic Thesaurus Discovery.* Kluwer International Series in Engineering and Computer Science, Vol 278.

Hearst, M. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of the fourteenth International Conference on Computational Linguistics*, Nantes, France, July 1992.

Hull, R. and Gomez, F. 1993. Inferring Heuristic Classification Hierarchies from Natural Language Input. Telematics and Informatics, 9(3/4), pp. 265-281.

IRS (Internal Revenue Service). 2001. *Tax Guide 2001*. Publication 17. http://www.irs.gov/pub/irs-pdf/p17.pdf

Lawrie, D., Croft, W. B., and Rosenberg, A. 2001. Finding topic words for hierarchical summarization. 24th ACM Intl. Conf. on Research and Development in Information Retrieval, 349-357, 2001.

Miller, G. (1995). WordNet: A Lexical Database for English. *Communications Of the Association For Computing Machinery (CACM)* 38, 39-41.

Sanderson, M. and Croft, B. 1995. Deriving concept hierarchies from text. Proceedings of the 22nd Annual Internationaql ACM SIGIR Conference on Research and Development in Information Retrieval, 160-170.

Sekine, S., Sudo, K. and Ogino, T. 1999. Statistical Matching of Two Ontologies. *Proceedings of ACL SIGLEX99 Workshop: Standardizing Lexical Resources*.

Zhang, K., Wang, J. T. L. and Shasha, D. 1996. On the Editing Distance between Undirected Acyclic Graphs and Related Problems. *International Journal of Foundations of Computer Science* 7, 43-58.