

Conceptual Browsing

Inderjeet Mani

703-983-6149 • imani@mitre.org

MITRE Sponsored Research



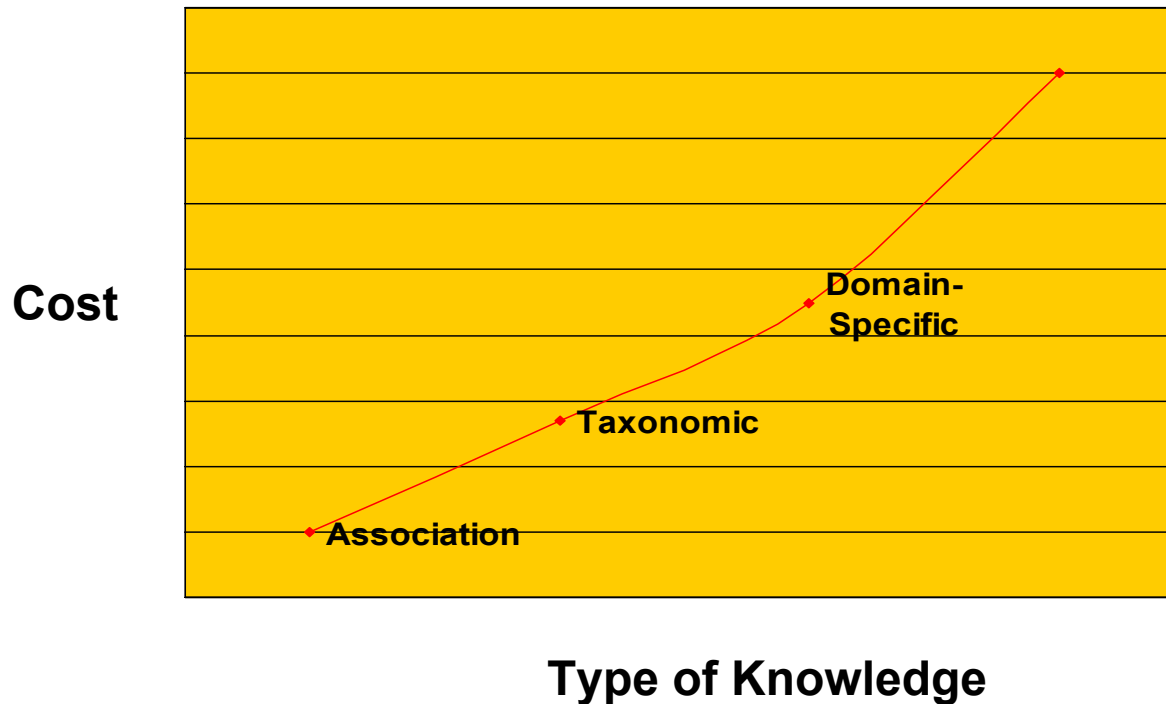
MITRE
Technology
Program

Problem

- **Humanity creates more than an exabyte (10^{18} bytes) of unique information each year.***
- **This information needs to be organized to help people satisfy their information needs.**
- **Ontologies are increasingly being used in applications like electronic commerce, bioinformatics, and corporate knowledge management.**
- **However, it is very expensive to build ontologies by hand, so methods are needed to help automate this.**
- * <http://www.sims.berkeley.edu/how-much-info/summary.html>

Background

- Knowledge engineering costs need to be reduced!



Objective

- **Automatically organize large quantities of unstructured data so that people can quickly and easily find the information they want**

Current Method

No ontologists? How d'you expect me to create an ontology for this domain? And who's going to pay for maintenance?



User creates/edits knowledge bases by hand

New Method

An exabyte of data? Wow! OK, mine the data to create the ontology automatically! Then let me edit it.....



System assembles domain-specific knowledge base, which user edits

Activities

- **Developed a domain-independent system for ontology induction**
 - **required: a collection of documents about a particular subject**
 - **applied to: collections on diseases (ProMed email), news (TREC), proteins (MedLine), tax information (IRS), technical books, etc.**
- **Integrated system with Veridian's ThemeLink search engine**
- **Evaluated ontology induction method**
- **Began the process of technology transitioning**

Highlight

Discovering Salient Terms

- Compare distribution of terms in domain corpus with distribution in background corpus
 - e.g., “income tax” is much more characteristic of IRS than of Reuters

Source	Documents with ‘income tax’	Documents without ‘income tax’	Total
IRS Publication 17	285 k_1	0	285 n_1
Reuters Corpus	9 k_2	19,024	19,043 n_2
Total	294	19,024	19,328

Log-likelihood Ratio

$$-2\log_2(H_0(p; k_1, n_1, k_2, n_2)/H_a(p_1, p_2; n_1, k_1, n_2, k_2))$$

Highlight

Inducing Relations

■ SubPhrase Relations

- ‘electric car’ is a *kind of* ‘car’
- ‘federal income tax’ is a *kind of* ‘income tax’

■ Existing Ontology Relations

- e.g., WordNet
 - ‘tailpipe’ is a *part of* ‘automobile’
 - ‘spouse’ is a *kind of* ‘person’

■ Explicit Patterns Relations

- ‘air toxics such as benzene’ => benzene is a *kind of* ‘air toxic’

■ Contextual Subsumption Relations

- ‘mosquito’ subsumes ‘mosquito pool’ and ‘standing water’

Evaluation: Term Weighting

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

TREC Topic 230 Term Percentile Rankings

Also, the system discovered 94% of (82) terms in hand-built list from IRS, and 58% of 1048 terms in a ProMed taxonomy produced by a bioterrorism analyst

Evaluation: Relation Discovery

- Each subject was asked to read four newspaper articles from TREC topic-230 sub-collection. Subject then answered 10 questions in each experiment, with articles remaining accessible in browser.
- In Experiment 1, the subject was asked to judge whether two terms were related in at least one of the documents; e.g., is “*Is horsepower related to electric car?*”
 - 5 questions chosen at random based on relations that weren’t labeled *kind-of* by system, 5 questions chosen at random based on pairs of terms not related by any relation
- In Experiment 2, the subject was asked to judge, based on 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?*”
 - 5 questions chosen at random based on relations that were labeled *kind-of* by system, 5 questions chosen at random based on pairs of terms not related by any relation

Relation Discovery Results

TREC Corpus

Is X related to Y?

System	Human	
	Related	Not-Related
Related	61 ✓	19
Not-Related	36	44 ✓

16 subjects
70 decisions involving Contextual Subsumption
10 involving WordNet

Kappa (3 subjects) = 0.53

Is X a kind of Y?

System	Human	
	Related by kind-of	Not-Related by kind-of
Related by kind-of	56 ✓	18
Not-Related by kind-of	6	74 ✓

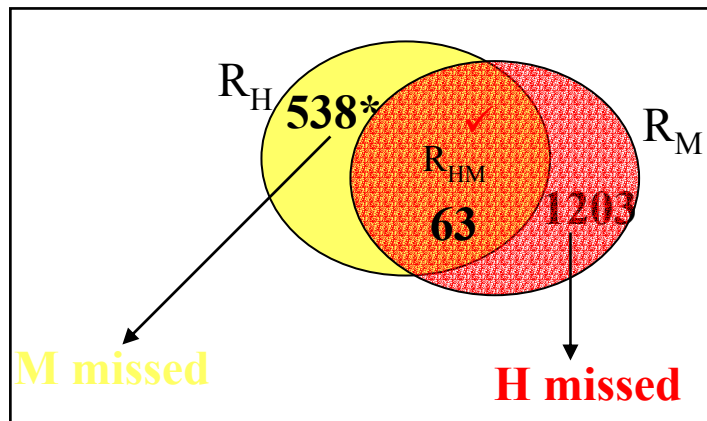
16 subjects
62 decisions involving SubPhrase Relations
10 involving WordNet

Kappa (3 subjects) = 0.72

Automatic Ontology Evaluation

- Desirable, since human experiments can't be used for anytime evaluation
- Difficult, because of
 - Lack of a unique gold standard to use as a reference ontology
 - Differences in terminology, structure, granularity
 - Differences in tasks for which the ontology was constructed
 - Comparing different versions of same ontology is relatively easy
- *Our approach:*
 - put aside the terminology issue
 - restrict term selection to terms in the human ontology
 - then compare machine and system relations closed within those terms

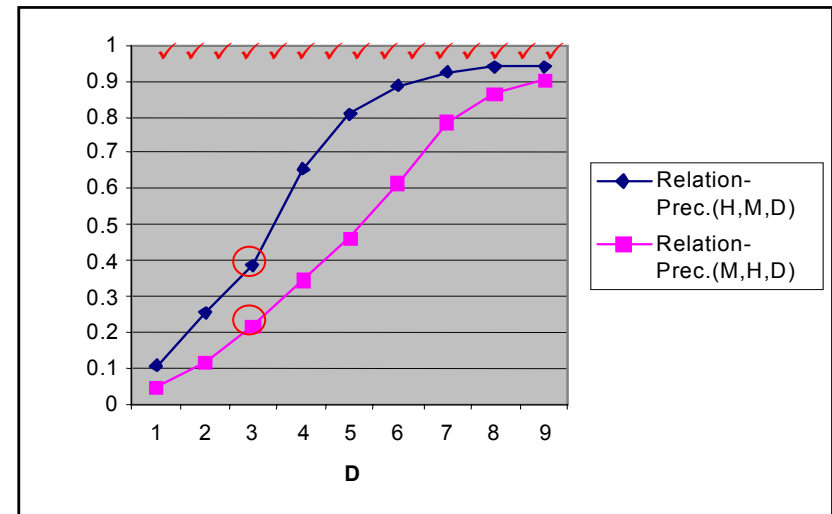
Results: Automatic Ontology Comparison



Relations Induced by H and M

- **M missed** example: 'maalox' is a *kind of* 'acid indigestion medicine'
- **H missed** example: 'chiropractic medicine' is a *kind of* 'medicine'
humans have substantial errors of omission!

- **Relation Precision (A, B, D)** = proportion of distance 1 relations in A that are at most a distance D apart in B
- 22% of distance 1 relations in M are at most 3 apart in H
- 40% of distance 1 relations in H are at most 3 apart in M



Impacts

- **USGC (building ontologies for IR): Using CBrowse v.1.02**
- **IRS (concept-based searching for customer service):**
 - **Demo part of IRS CCOF (Contact Center of the Future) - joint work with Margot Peet's MSR**
- **NSF (building ontologies for bioinformatics):**
 - **MITRE collaborating with Georgetown on NSF ITR Program "Protein Ontology Induction"**

Future Plans

- **Task-based evaluation**
 - **test effectiveness of induced ontology in query expansion**
- **Add more knowledge sources**
- **Further work on bioinformatics domains**
- **Further integration with IRS CCOF**