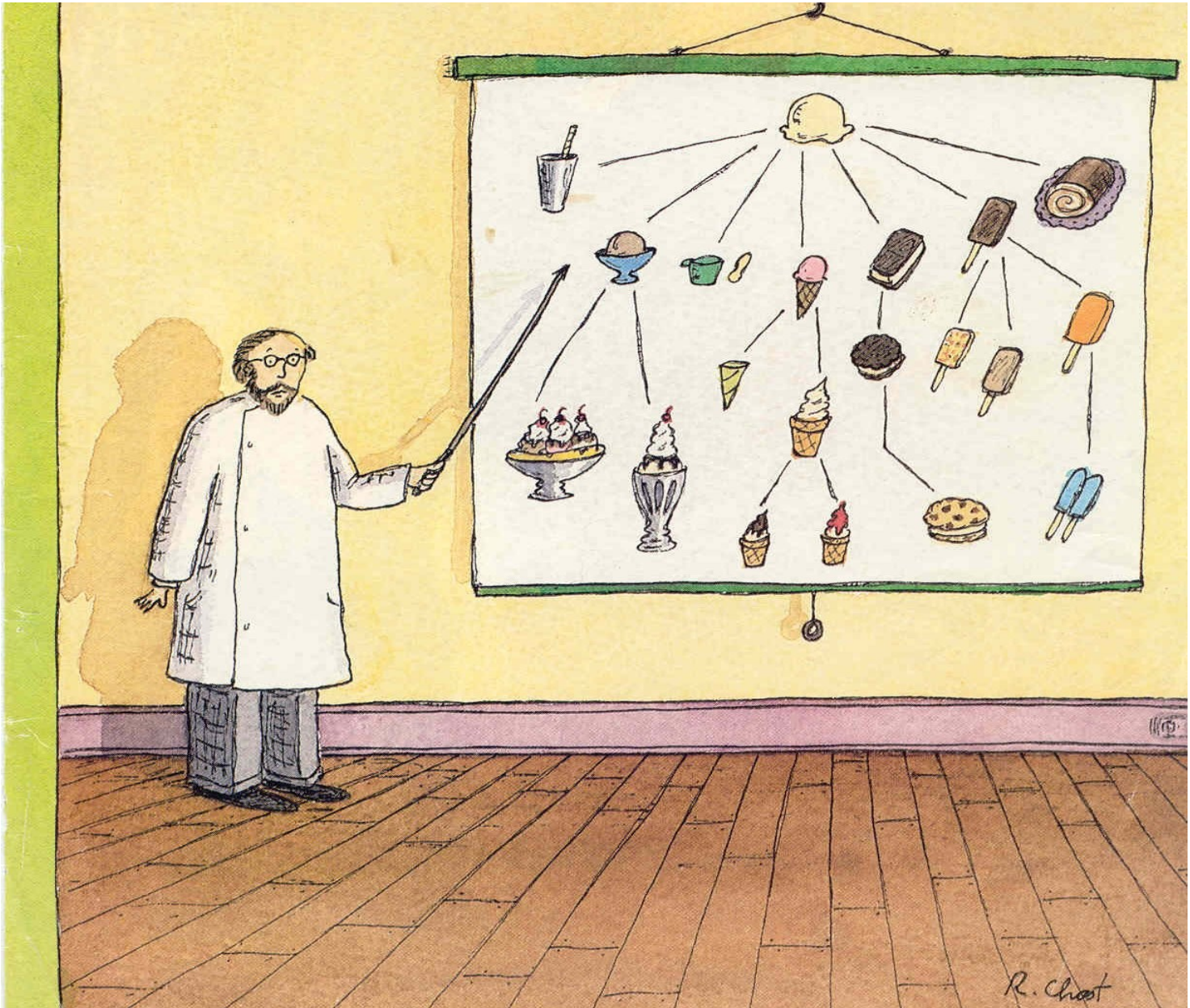


Ontology Learning and Population from Natural Language Corpora



Mathias Niepert
Spring 2006

What is an Ontology?

- Abstract, simplified view of the world (domain) that we wish to represent
- Formal specification of a conceptualization (Gruber 1993)
- Most common relations between concepts:
Taxonomic and non-taxonomic
- Taxonomic: “is-a”, “is-part-of”, ...
- Non-taxonomic: “teacher-of”, “agrees-with”, ...

Why develop an Ontology?

- To analyze and enable reuse of domain knowledge
- To share common understanding of the structure of information
- To separate domain knowledge from operational knowledge
- To improve the representation and accessibility (searching, cross-referencing, clustering, ...) of documents

Learning and Populating an Ontology

- In the past: mostly by hand (bottleneck)
- Goal: semi-automatic, iterative
- 4 major problems:
 - Finding the classes (concepts)
 - Finding taxonomic relations between classes
 - Finding non-taxonomic relations between classes and instances
 - Evaluation!

1. Problem: Finding classes (concepts)

- Information extraction techniques (e.g., name detection; “George Berkeley”), N-grams (e.g., “Philosophy of Mind”), TFIDF, ...
- Existing Ontologies (e.g., WordNet, ...)
- Structure of documents (e.g., headers, titles, links,...)
- Use cluster of synonyms as concept (e.g., “doctrine”, “philosophy”, “school of thought”)
- Important: Normalization (e.g., only use the singular form for all concepts; generally improves recall)

2. Problem: Learning taxonomic relations

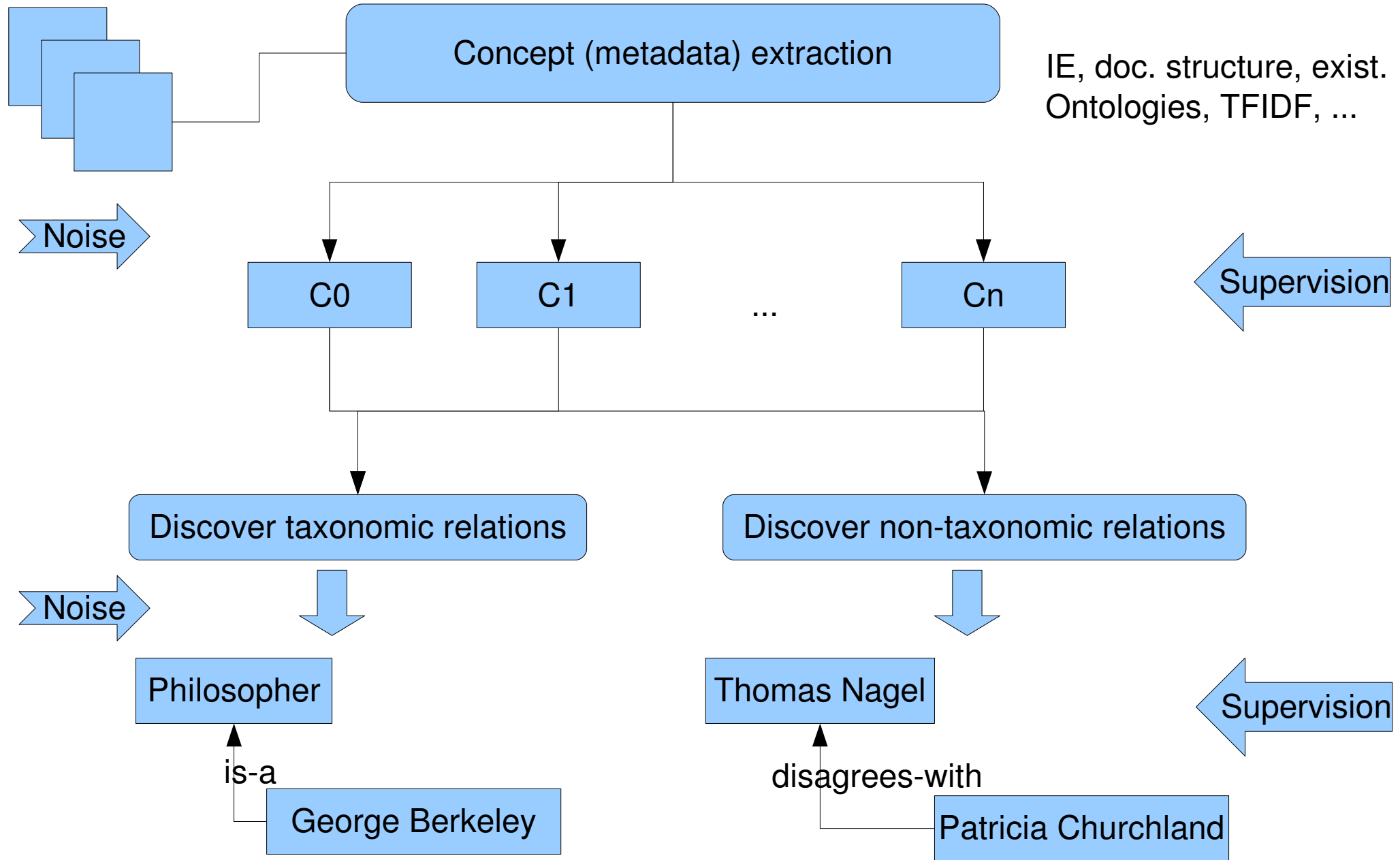
- Clustering techniques (hierarchical, ...)
- Hearst Patterns (e.g., NP_0 such as NP_1, NP_2, and I or NP_n) (Hearst 1992)
- Existing Ontologies (e.g. WordNet, Cimiano 2004) & classification, difficult in specialized domains
- Other heuristics, e.g. “international meeting”, “meeting” (Velardi 2001)
- World wide web, Cimiano et al.: Use Google API <Ci>s such as <Cj>, <Ci> and other <Cj>s, ...

3. Problem: Non-taxonomic relations

- “Frequent Itemset Mining” (Kavalec 2005)
 - Use units of text as windows (one sentence, 8 terms, ...)
(what is the optimum?)
 - Only use certain types of sentences
 - If two or more of the concepts appear within the window
——▶create itemset
(can weighting policies improve accuracy?)
 - Save the verb to specify the kind of relation
 - Use support, confidence and/or other measures

Ontology Learning and Population from Natural Language Corpora

documents



- Essentially, lots of NLP problems appear in all three sub-problems!
- They all introduce noise and must be supervised by domain specialists
- Evaluation is very difficult (Existing Ontologies, Domain Experts)
- Our “collaborative” project...

The Stanford Encyclopedia of Philosophy¹

- First “Dynamic reference work”
- Collaborative writing, publishing and maintaining
- More than 1000 authors and subject editors
- New articles and updates refereed by one or more of the nearly 100 editors on the editorial board before publication
- More than 2,500,000 entries downloaded each month

¹ <http://plato.stanford.edu>

The SEP Project

- Colin Allen, editor and principal programmer of the Stanford Encyclopedia
- One student from Philosophy, one from CS
- Goals:
 - Build a robust Ontology
 - Adjust and update this Ontology semi-automatically
 - Based on this Ontology, implement amazing search and cross-referencing engine
 - Evaluation of Ontology by domain experts “on the fly”

The SEP Project

- Real world data, very broad domain
- Algorithms will be implemented in the SEP and heavily used
- Takes advantage of the authors' feedback for
 - Evaluation of the employed methods
 - Adjustment of algorithms
- Could take advantage of user search traces (future work)

The SEP Project

- Authors' feedback:
 - Feedback on automatically generated concepts/instances/relations (“Kant criticized Metaphysics?” y/n)
 - New concepts/instances/relations but also cross-references, which can be used to improve the Ontology (iterative process)

References

- Camino et al., Learning Taxonomic Relations from Heterogenous Evidence, 2004
- Bisson et al., Designing clustering methods for ontology building, 2000
- Haase, Völker, Ontology Learning and Reasoning, 2005
- Kavalec, Svátek, A Study on Automated Relation Labelling in Ontology Learning, 2005
- F. Noy, McGuinness, Ontology Development 101