Learning Domain-Specific Ontology from Scientific Publications for Knowledge Management

Abstract

1 Introduction

Over the last decade, ontology engineering has received considerable attention in the AI community, accelerated by the Semantic Web movement (). Automatic or semi-automatic ontology induction and ontology acquisition methodologies that utilize machine learning and natural language processing techniques to discover ontological knowledge from various data resources have been leading the research trend.

In this paper we describe a weakly-supervised construction of a light-weight ontology for the computational linguistics research community, based on a corpus of papers published during the last few decades. This task can be categorized to automatic ontology population, or Ontology Driven Information Extraction (Bontcheva and Cunningham, 2003), whose goal is extracting and classifying instances of concepts and relations predefined in an ontology. The motivation of developing an ontology for scientific research is originated from the increasing interest of studying the dynamics of reasearch communities.

This is in contrast to standard upper ontologies that incorporate philosophical understanding of concepts' categories and designed to be applicable to general tasks, such as SUMO (Suggested Upper Merged Ontology) and Cyc upper ontology (OpenCyc). Our ontology structure is also different to the various domain ontologies and task ontologies that already exist. (How?...)

Although hierarchical struture is commonly used in standard ontologies, to better serve the application purposes, we define the ontology in the format of a graph, each vertex representing a concept , with a required class tag attaching to it, representing the concept class it belongs to.

Concept Classes

We define the following classes in the ontology: *Task(T), Method/Algorithm(M), Software/Tool(S), Language(L), Corpus(C), Other(O)*

Relations

There are two basic types of relations: intra-class relation and inter-class relation.

Inter-class relations are undirectional, and the semantic interpretation is uniquely decided by the classes of the two concepts. For example, if the concept "Part-Of-Speech tagging" of class *T* is connected to the concept "Hidden Markov Model" of class *M*, there should be a unique interpretation "*isUsedFor* (Hidden Markov Model, Part-Of-Speech tagging)". Except for the *Language* class, whose relations are constrained to *T*, *S* or *C*, every possible pair of classes have a uniquely defined relation. (add definitions to appendix)

Intra-class relations are directional and defined only within class *T*. There are two kinds of relations between two Task concepts: "*ISA*" and "*isUsedIn*". "*ISA* (A, B)" relation describes task A as a special case of task B, or a subfield of B. For example, we can have *ISA* (Statistical Machine Translation, Machine Translation). *isUsedIn*(A, B) relation describes task A as a sub-component of task B, which means A is performed to achieve B. Similar to the *isUsedFor* (M, T) relation, This is a "soft" link since usually a path to approach a problem is highly variable among researchers.

2 Approach

2.1 Terminology Extraction

2.2 Relation Extraction

2.3 Concept Class Classification

2.4 Validation

Uncertainties and controversies sometimes arise during the

3 Experiment

3.1 Data

3.2 Baseline

4 Evaluation

The ontology graph constructed using the above methods contains [ ] vertices (concepts) and [ ] links (relations). Since we are lack of a gold standard existing ontology that encodes all relations we are interested in, we adopt a data-driven appraoch, described by (Brewster et al., 2004). We sample the results by taking a part of the ontology that is around a specific concept out and evaluate its fitness to that particular domain of knowledge. We picked [ ] well-studied NLP tasks (enumerate...) and selected for each topic a comprehensive survey article (the surveys are not in the corpus we used to build the ontology). An ideal ontology should encode every concept and relation of interests that appear in the survey article, and also free of false relations or irrelevant concepts. The evaluation is seperately done for the two aspects: concepts and relations.

4.1 Concept Level Evaluation

extracted a set of relevant domain-specific terms from the corpus of documents, using latent semantic analysis. The amount of overlap between the domain-specific terms and the terms appearing in the ontology (e.g. as names of concepts) can then be used to measure the fit between the ontology and the corpu

4.2 Relation Level Evaluation

To create an accurate and complete gold standard for relations, we extracted all sentences in the survey article that contain at least two concepts in the ontology or one concept and a pronoun (considering correfence) and manually extracted the relations of interests. In calculating the precision, we only consider the relations between concepts that both are validated by the survey (overlapping concepts in the concept level evaluation).

5 Related Work

5.1 Terminology Recognition

5.2 Relation Extraction

(Hearst, 1998) (Brin, 1999)

5.3 Automatic Ontology Population

6 Conclusion

References

Kalina Bontcheva and Hamish Cunningham, 2003. *The Semantic Web: A New Opportunity and Challenge for Human Language Technology*, page 8 p. Citeseer.

C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks. 2004. Data driven ontology evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal.

Sergey Brin. 1999. Extracting patterns and relations from the world wide web. In *Selected papers from the International Workshop on The World Wide Web and Databases*, pages 172–183, London, UK. Springer-Verlag.

Marti Hearst. 1998. Automated discovery of wordnet relations. In Christiane Fellbaum, editor, *An Electronic Lexical Database and Some of Its Applications*, pages 131–151, Cambridge, MA. MIT Press.