Unsupervised Language Learning: Semantic Parsing and Beyond

Hoifung Poon **Pedro Domingos**
Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195-2350, U.S.A.
{hoifung,pedrod}@cs.washington.edu

1 Introduction

Language learning, natural or artificial, is a deep and fascinating challenge. In the past, there have been considerable successes in various subtasks such as syntactic parsing, but there has been a lack of concerted efforts in learning a complete end-to-end system for language understanding. Recently, there has been increasing interest in developing new machine learning algorithms for both representing complexity in problem structures and handling uncertainty [5, 1, 2]. Building on these advances, we propose to tackle natural language understanding by incorporating the various subtasks and a small amount of prior knowledge into a coherent model, and leveraging large-scale of joint inference to enable effective learning with little or no labeled data.

We have taken some initial steps in this direction. For example, we built an unsupervised coreference resolution system that rivals the performance of state-of-the-art supervised approaches by leveraging joint inference and prior knowledge [7]. Most recently, we developed the USP system, which is the first approach for unsupervised semantic parsing [8].

Semantic parsing aims to obtain a complete canonical meaning representation for input sentences. Past approaches either manually construct a grammar or require example sentences with meaning annotation, and do not scale beyond restricted domains. Unlike previous works on semantic parsing, USP does not predefine a formal meaning representation, but rather discovers its own internal representation during learning. The key idea of USP is to recursively combine subexpressions that are composed with or by similary subexpressions. USP is defined in just four formulas in Markov logic [3]. The core of USP is a joint probability model for identifying subexpressions that are meaning units and resolving the variations among them for the same meaning. USP inputs dependency trees of sentences, and outputs the learned parser as well as an extracted knowledge base formed by the MAP semantic parses. We evaluated USP by applying it to extract knowledge from biomedical abstracts and answer questions. Compared to the previous state-of-the-art unsupervised systems, USP obtained the highest precision of 88%, and extracted three times of correct answers as the second best [8].

USP only induces a flat level of clusters without any subsumption hierarchy among them. The next major step forward is to induce a full-fledged ontology from text. Ontology induction is not only an important goal in its own right, but also promises to bring about exponential amount of speed-up in learning by enabling the use of coarse-to-fine inference [4, 6]. The key idea is to start with a flat ontology, and recursively induce higher-level clusters to summarize similar aspects among existing ones while learning the semantic parser. The output is a semantic parser, an ontology, and the MAP parses for input text. In effect, this jointly conducts ontology induction, population, and knowledge extraction. These three tasks are only separately pursued in previous work. Other major directions include incorporating additional subtasks such as dependency parsing into the joint model, as well as modeling other phenomena of language learning such as quantification and discourse processing.

References

[1] G. Bakir, T. Hofmann, B. B. Schölkopf, A. Smola, B. Taskar, S. Vishwanathan, and (eds.). *Predicting Structured Data*. MIT Press, Cambridge, MA, 2007.

[2] Yoshua Bengio. Learning deep architectures for AI. In *Foundations and Trends in Machine Learning, 2:1*, 2009.

[3] Pedro Domingos and Daniel Lowd. *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan & Claypool, San Rafael, CA, 2009.

[4] Pedro F. Felzenszwalb and David McAllester. The generalized A* architecture. *Journal of Artificial Intelligence Research*, 29, 2007.

[5] Lise Getoor and Ben Taskar, editors. *Introduction to Statistical Relational Learning*. MIT Press, Cambridge, MA, 2007.

[6] Slav Petrov. *Coarse-to-Fine Natural Language Processing*. PhD thesis, Department of Computer Science, University of Berkeley, 2009.

[7] Hoifung Poon and Pedro Domingos. Joint unsupervised coreference resolution with Markov logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 649–658, Honolulu, HI, 2008. ACL.

[8] Hoifung Poon and Pedro Domingos. Unsupervised semantic parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, Singapore, 2009. ACL.

Topic: learning algorithms, speech and auditory processing

Preference: oral