A Framework for Representing and Jointly Reasoning over Linguistic and Non-linguistic Knowledge

by

Arthi Murugesan

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the degree of

DOCTOR OF PHILOSOPHY

Major Subject: Cognitive Science

The original of the complete thesis is on file
In the Rensselaer Polytechnic Institute Library

Examining Committee:

Nicholas L. Cassimatis, Thesis Adviser

Paul Bello, Member

Selmer Bringsjord, Member

Mark Changizi, Member

Ron Sun, Member

Rensselaer Polytechnic Institute
Troy, New York

October, 2009
(For Graduation December 2009)

ABSTRACT

Natural language poses several challenges to developing computational systems for modeling it. Natural language is not a precise problem but is rather ridden with a number of uncertainties in the form of either alternate words or interpretations. Furthermore, natural language is a generative system where the problem size is potentially infinite. This combination of uncertainty and generativity challenges existing computational solutions.

In order to model natural language, innovation is required at the fundamental level of computational algorithms. Existing algorithms are designed either to handle uncertainty or to handle generativity, not both. For example, Bayesian networks treat uncertainty in a generic mathematically well-founded fashion, but are however restricted to a fixed size and configuration during inference. On the other hand, conventional PCFG parsers like Earley parsers are not restricted to a particular finite set of possible words, but are incapable of handling other uncertainties as those brought up by the semantics of sentences.

Our approach to providing a computational framework capable of modeling natural language understanding is two-fold. Our first aim is to propose a formalism that is capable of representing various kinds of uncertainties, for example word sense ambiguity in syntax and uncertainty in referent resolution in semantics, in the same generic fashion. To follow up, we propose a working inference algorithm that overcomes the problems with large (often infinite) problem spaces entailed by such mechanisms.

Specifically, this thesis proposes a formalism, generative satisfiability language (GenSAT), capable of representing constraints over potentially infinite domains. An inference algorithm, GenDPLL, is then defined which is guaranteed for a well-defined sub-class (increasing cost theory, relevant models) of the possible problems expressed in GenSAT. In order to show that GenSAT and GenDPLL are capable of representing natural language, as a proposed first step, this thesis demonstrates how Probabilistic Context Free Grammar (PCFG) can be represented in GenSAT.

However, GenSAT is a fairly low-level representation, making the representation of PCFG constraints directly in GenSAT difficult to comprehend. Therefore, we have defined an intermediate language, generative probabilistic language (GenProb) that

abstracts over the details of GenSAT in a level of abstraction close to the conditional table representation in Bayes nets. GenProb is interesting in its own right, because it is well-defined with intuitive semantics and helps show case several properties of the problem encoded, like the property of increasing cost. The thesis then elaborately details the translation of a problem encoded in GenProb theory to a corresponding problem in the GenSAT language. The final chapter describes in detail how a PCFG problem can be represented in GenProb.

Hence by representing a linguistic grammar in the proposed formalism, GenProb which can be translated to GenSAT, and defining an inference algorithm GenDPLL for this language, we are able to show that language can be parsed using algorithms that unify syntax and semantics.