IITB-Monash Research Academy

Indian Institute of Technology Bombay | MONASH University

An Indian-Australian research partnership

Project Title:
"Learning for 'Deep' Semantic Parsing in Natural Languages by Integrating Knowledge-based and Statistical Techniques"

Project Number: IMURA0195

Monash Supervisor(s) : Prof. Geoff Webb

IITB Supervisor(s) :Prof. Pushpak Bhattacharyya, Prof. Ganesh Ramakrishnan,

## Project aims

Semantic parsing is the process of mapping natural language input to some formal representation of its meaning. Universal Networking Language (UNL) is a meaning representation language developed at the UNDL foundation in Geneva, Switzerland. Being the pivot language in Interlingua-based systems, this is suitable for manipulation by a machine for automated reasoning. Our expertise in UNL could be used to leverage the semantic representation aspect of the problem being defined.

The problem of learning for semantic parsing can be seen as having two aspects to it:

- Learning semantic parsing requires knowledge of world and the linguistic rules.
- Learning the knowledge of the world (and also linguistic rules) requires 'understanding' of natural language (which can be done by semantic parsing).

There is a recursive dependency between the two tasks. Bootstrapping gives us a methodology to handle such problems. The idea is to start with a small knowledge-base and linguistic rules. Use these on the raw data (corpus of web-text) to learn more rules and get more world-knowledge. Iterate on the process to refine the knowledge and thereby enabling semantic learning.

The two main challenges one would encounter in tackling the issues mentioned above are:

- Noisy data: Essentially the input to the system as well as the data available for training would be incomplete, inaccurate and ambiguous.
- Complexity of the system being built: The system should not be over-engineered as it would be biased to the training data available. Ideally, even domain specific applications should be as simple and generic as possible in design.

A general framework to do this is provided by the relational learning techniques of machine learning. Inductive logic programming (ILP) is a relational learning approach which generalizes from individual instances / observations in the presence of background knowledge. Statistical relational learning (SRL) attempts to represent reason and learn in domains with complex relational and rich probabilistic structure.

The intuition motivating the application of the above techniques is as follows:

- SRL: Would help us to grapple with the issue of noisy data. It provides a probabilistic framework to tackle inaccuracy and ambiguity in natural language text.
- ILP: Would help us to provide the linguistic background knowledge based on our understanding of the domain. It would be more intuitive to deal with rules in defining the system.

SRL in combination with ILP provides us the tools to model uncertainties in natural language in the framework of a solid knowledge-base. An investigation in this direction to solve the semantic parsing problem, might give the research community new insights and also advance the state-of-the-art.