

CS600-Level Project Proposal

Student Name: Philip Nadeau

Student ID: W00841188

E-mail: nadeaup@cc.wvu.edu

Starting Date: 1-Apr-2008

Supervisor: Dr. Jianna J. Zhang

E-Mail: Jianna.Zhang@wvu.edu

Project Title: Self-Improving Message Understanding Systems with Genetic Programming

Project Description:

Information Extraction is the term for extracting meaning from textual data such as public records and newspapers. The goal of a *Message Understanding System* (MUS) is to automatically build object models of events and actors from this data so that it may be mechanically queried. Linguists would refer to this object model as a representation of the *semantics* (meaning) of the documents. Because of this object model, MUS are more powerful than keyword-based retrieval systems, which are semantically naïve.

At least one successful MUS, FASTUS¹, was built using a cascading arrangement of *finite state transducers*, which are automata that have separate input and output streams. FASTUS used these automata to implement *partial parsing*, which extracted information based on common sentence templates in sources of interest.

The following chain of reasoning illustrates the key intuition:

- Information extraction is essentially the mechanical extraction of a semantic model from text.
- The figure of merit for an information extractor is how much correct information it extracts from a source, relative to the amount of information that can possibly be extracted.
- The sources and the extractable facts comprise a set of training examples, and therefore information extraction may properly be regarded as a machine learning problem.
- Genetic algorithms have been used to induce context-free grammars for natural languages. Genetic algorithms (and specifically genetic programming) should therefore be usable to induce a FASTUS-type cascading-automata MUS.
- *Therefore, it is possible to build a self-improving message understanding system based on genetic programming.*

The goal of this project is to evaluate the feasibility of genetically self-improving MUS, first through a survey of applicable literature, and then by construction of a prototype system over a limited problem domain (for example, weather reports.) The idea is believed to be novel and it is also believed that there are few (if any) published results on this particular synthesis of methods.

Planned Paper:

The project shall be conducted as three major tasks, corresponding to each of the classes in the 601/602/603 sequence. The first task is to conduct a literature survey to identify sources where genetic algorithms were applied to information extraction. The goal of this task is to produce a survey suitable for distribution within the department.

¹ FASTUS is a product of the Message Understanding Conferences (MUC) that were conducted in the 1990's by SAIC on behalf of DARPA. http://www-nlpir.nist.gov/related_projects/muc/

The second task is to design an extraction system and an experiment to test it. In order to keep the task within the scope of available resources, the system will be restricted to a simple subset of semantic information (for example, weather conditions) and a suitable corpus will be identified (for example, a wire service archive.) Unless a better model is found during the survey, the experiment will be based on cascading automata of the FASTUS type, because it has been proven in practice.

The third task is to execute the experiment. The goal is a MUS that can correctly extract the majority of information from a corpus and answer simple queries (ex: Show all locations where hailstorms occurred between these dates.) The results will be recorded and documented to establish a performance comparison with other MUS discovered during the survey.

The results of all three tasks would be synthesized into a paper for submission to a conference or journal dealing with machine learning, genetic algorithms, natural language processing, or information extraction.

Time Line:

1. Literature Survey. (601, Spring 2008)
 - a) Identify sources where genetic algorithms produce grammars or extract semantics.
 - b) Identify milestones in Message Understanding Systems.
 - c) Obtain documents.
 - d) Source summary and prioritization, Pass 1.
 - e) Detailed source summary, Pass 2.
 - f) Rough Draft.
 - g) Preliminary design of a usable system.
 - h) Final Draft.
 - i) Distribute draft to interested parties within the Department.
2. Summer 2008
 - a) Identify and obtain a pre-parsed corpus for use as training examples.
 - b) Design a mechanically searchable object model for the semantics of the corpus.
 - c) Find a suitable metric for measuring the performance of information extractors.
 - d) Design a genetic algorithm that uses the metric to evolve cascading networks of automata.
3. Implementation and Experiment (602, Fall 2008)
 - a) Implement the designed algorithm.
 - b) Execute system and measure results. Record and document all results.
 - c) Identify journals and conferences for submission.
4. Produce paper for publication in a journal or conference proceedings. (603, Winter 2009)