

Logic Programming for Natural Language Processing

Menyoung Lee

October 3, 2005

1 Problem Statement

Natural language processing is concerned with the problem of parsing and understanding documents written in a natural, human language.

2 Purpose, goals, and value

The main goal of the project is to develop a method and system for posing queries about information contained in texts belonging within a particular domain of knowledge. Hence, the first goal is the creation of a GATE information extraction application to read plain English documents in a particular subject domain, and process the information contained in them into Prolog facts. Then, these facts, along with a number of subject matter expert (SME) rules can be used by a Prolog system to answer queries based on fact and logic, and therefore the second phase goal is to program an expert shell that draws from extracted knowledge.

Current information retrieval technologies (such as Google Search) can only provide relevant documents, which still must be read by the subject matter expert for proper extraction of facts and interpretation. Thanks to the Internet, information in electronic texts are available in quantities that no human can possibly hope to process, meaning that much information is missed. Therefore the desire to recoup some of the opportunity cost of missed information has spurred research in text understanding.[1] Processing extracted information into Prolog facts can add the power of deductive logic to our current text analysis capabilities, bringing Artificial Intelligence one step closer to the dream of replacing the human expert.

3 Scope of Study

The documents to be processed will be within the subject area of the history of mathematics, the main corpus of them being biographies of 1996 mathematicians from the MacTutor History of Mathematics archive at <http://www-history.mcs.st-andrews.ac.uk/>

4 Background and Review of Literature

Broadly, information extraction refers to any method for parsing information from a large corpus of text, including the retrieval of appropriate documents and the tagging of terms. More narrowly, information extraction is the process of identifying instances of a certain class of objects and extracting the relevant attributes and relationships.[6] One very early development in IE was Zarri's RESEDA, a semantic metalanguage used to describe various French historical figures and the relationships between them.[9] Another much more recent work quite similar to my proposed project sought to construct a knowledge base from faculty websites by applying machine learning algorithms to train an IE system.[2] A Prolog-based IE system in the field of bio-genetics is presented in [5]. Current research is almost entirely focused on IE tasks within a single domain with a human-made ontology, a limitation that may be a little foolhardy of me to try to break.

Two specific technologies will be very important to this project: The General Architecture for Text Engineering (GATE) and the XSB System (XSB). GATE provides a comprehensive architecture for more easily developing natural language processing applications, and its Java Annotation Patterns Engine (JAPE) provides the capability to add custom grammars to control how documents are annotated.[3] [4] XSB at the core is a Prolog system with its full capability of executing first-order predicate logic. In addition it is a massive improvement over typical Prolog through the implementation of tabling, also known as memoizing, which means the Prolog system does not recompute a repeated query from scratch. Its ability to act as an efficient deductive database engine for storing extracted information. [7]

5 Procedure

1. Create an ontology appropriate for the subject area, i.e. history of mathematics. As far as I can tell, this is the part where current state of the art isn't even close to being able to generate automatically from text. Oil-Ed seems to be an

appropriate ontology editor for creation of an ontology in DAML format.

2. Annotate the text corpus through GATE and its Nearly New Information Extraction (ANNIE) system with tags such as Parts of Speech tags.
3. Develop an ontology-aware Gazetteer, in the Java Annotation Patterns Engine (JAPE), to identify the appropriate instances of the classes specified in the math-history ontology.
4. Program an application in Prolog to read the marked up output and process them into facts and compiling them into a database. XSB System is a Prolog system that does memoization to optimize its performance, and it can be used here. (e.g. "Socrates is a man" becomes `is_a('Socrates', man)`)
5. Add SME rules to the Prolog program, thus making it a complete expert shell. (e.g. `is(man, mortal)` is added to our Prolog application.)
6. Develop an interface for the posing of queries to this expert shell. Interprolog makes a nice Java front-end for XSB. (e.g. XSB answers "yes" to `-? is('Socrates', mortal)` because it can be deduced from the extracted information and the SME rule.)
7. Continue to improve upon the shell by training the ANNIE pipeline and possibly applying some learning algorithms as in some studies [2] [8]
8. Measurables: the proportion of queries able to be correctly answered, some time/memory efficiency benchmarking of the databasing and reasoning.

References

- [1] COWIE, J., AND LEHNERT, W. Information extraction. *Communications of the ACM* 39 (1996), 80.
- [2] CRAVEN, M., DIPASQUO, D., FREITAG, D., MCCALLUM, A. K., MITCHELL, T. M., NIGAM, K., AND SLATTERY, S. Learning to extract symbolic knowledge from the World Wide Web. In *Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence* (Madison, US, 1998), AAAI Press, Menlo Park, US, pp. 509–516.

- [3] CUNNINGHAM, H., MAYNARD, D., BONTCHEVA, K., AND TABLAN, V. Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (2002).
- [4] CUNNINGHAM, H., MAYNARD, D., AND TABLAN, V. Jape: a java annotation patterns engine (second edition). Research Memorandum CS-00-10, Department of Computer Science, University of Sheffield, Nov 2000.
- [5] EMMS, M. A prolog based information extraction system. In *Proceedings of the International Applications of Prolog Conference* (2001), pp. 160–167.
- [6] GRISHMAN, R. Information extraction: Techniques and challenges. In *Proceedings of the Summer Convention on Information Extraction* (1997), pp. 10–27.
- [7] SAGONAS, K., SWIFT, T., AND WARREN, D. S. XSB as an efficient deductive database engine. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (1994), pp. 442–453.
- [8] SCHNEIDER, R. A lexically-intensive algorithm for domain-specific knowledge acquisition. In *New Methods in Language Processing and Computational Natural Language Learning* (1998), pp. 19–28.
- [9] ZARRI, G. P. Automatic representation of the semantic relationships corresponding to a french surface expression. In *Proceedings of the Conference on Applied Natural Language Processing* (Santa Monica, CA, 1983), pp. 143–147.