Overview

We propose a bifurcated paradigm for the construction of a Prolog knowledge base from a body of documents: first, an information extraction (IE) application that will annotate the corpus and output the annotated documents, and second, a Prolog knowledge base (KB) application that will transform the annotated documents into a KB (a set of facts). The General Architecture for Text Engineering (GATE) was used as the platform for the development and execution of the IE application, which included most components of A Nearly New Information Extraction (ANNIE) system. Apart from the basic IE capabilities of ANNIE, the application featured additional high-level annotation grammars written in the Java Annotation Patterns Engine (JAPE) language and a trainable annotator that used the maximum entropy machine learning model, which were designed to annotate several biographies of well-known mathematicians. The Prolog KB application, programmed to be executed within the XSB System, was designed to receive the annotated text output by the IE application and produce a knowledge base, and it successfully creates a database of Prolog facts that can be intelligently queried through the XSB System. The KB utilizes the frame representation of facts, specifically by treating one document as an object to be represented as a frame, with each annotation type treated as a slot whose multiple values are whichever specific strings were annotated by the IE application. This transformation of extracted information into Prolog facts is a link between IE, a recent development in Natural Language Processing, and logic programming with Prolog.

Background

Natural Language Processing (NLP) is the automated understanding and generation of text written in a natural language such as English. While the ill-defined notion of complete text “understanding” is far beyond the grasp of current state of research, intelligent systems with the ability to automate at least some of the tasks in understanding text can be of assistance to a human expert in reading and analyzing a large corpus of documents. The advent of the Internet and the subsequent explosion in the sheer volume of textual information readily available in electronic form presents new opportunities for exploitation of the available information, while simultaneously it presents a challenge, as it is impossible for an analyst to read so much text. Hence arose Information Extraction (IE), a subset of NLP that calls for the transformation of information contained in free, unstructured text into a prescribed structure, specifically by identifying instances of certain objects, their attributes, and/or relationships between them. The purpose of this project is to link this new development in NLP with classical logic programming by developing an application that construct a Prolog knowledge base of information extracted from text.