Automating Information Extraction

Elder David W. Embley

Abstract

Problem: Efficiently harvest and organize family-history information from the vast and growing collection printed documents (mostly scanned and OCRed historical documents).

Motivation: Enhance patron search over extracted information, contributing to more enthusiasm about and success in finding ancestors and preparing names for temple work.

Solution: Automate extraction (see below).

Evaluation: Measure accuracy and cost.

Production: If the accuracy is deemed high enough while the cost is deemed low enough, the research project should be turned into a development project.

Executive Summary of Projects

The church has millions of printed documents containing valuable family-history information—already millions of scanned and OCRed page images from family history books and newsletters and thousands of born-digital obituaries and no end in sight to what will likely be collected. Extracting genealogy information from this document collection is a daunting task: its volume and variety and the velocity at which it is being collected pose problems beyond which we can address with current crowd-sourcing techniques.

* Ground Truthing Tool: We have created a form-based annotation tool. Users can click on text in document images to rapidly extract information. The church used the tool a year or so ago in a pilot project to hand-extract information from 300 or so randomly chosen books in the scanned-book collection. The objective of the pilot project was to estimate the overlap between information found in the collection with information found in the FamilySearch database and estimate the added value of collecting all the information. Results were promising. Since the time of this initial pilot project, the tool has be enhanced to permit more sophisticated relationships among extracted entities (while keeping the same simple form-filling motif for users), and a back-end evaluation tool for precision and recall has been added so that the tool can be used to establish the ground truth for research purposes to evaluate automated extraction tools.
* FROntIER: The acronym FROntIER stands for **F**act **R**ecognizer for **Ont**ologies with **I**nference and **E**ntity **R**esolution. It is a framework consisting of several tools to recognize and extract assertions stated in semi-structured text using an “extraction ontology” and to organize these assertions with respect to an ontological view. Organization includes, in addition to populating an ontology with these stated assertions, also populating the ontology with implied assertions using inference rules and resolving entity identity to remove duplicates. (For more details about FROntIER, see Attachment [1].) Implementation is complete, and we are just now beginning to evaluate the tool with respect to ground truth to be established for about fifty three-page sequences randomly selected from the semi-structured documents in the scanned book collection. We also plan to evaluate the tool for a collection of obituaries. The evaluation should provide cost/benefit data both for hand-annotation and for FROntIER-annotation. Although the cost/benefit analysis may show that knowledge-engineering with FROntIER is better than hand-annotation, we believe that it will still be too costly to apply in production unless the work of knowledge engineering can be significantly reduced. Hence, we have initiated three projects: PatternReader, OntoSoar, and GreenFIELDS.
* PatternReader: A flexible, accurate, and efficient method of extracting facts from lists in OCRed documents and inserting them into an ontology would help make those facts machine queryable, linkable, and editable. But, to work well, such a process must be adaptable to variations in list format, tolerant of OCR errors, and careful in its selection of human guidance. We propose a wrapper-induction solution for information extraction that is specialized for lists in OCRed documents. In this approach, we induce a grammar or model that can infer list structure and field labels in sequences of words in text. Second, we decrease the cost and improve the accuracy of this induction process using semi-supervised machine learning and active learning, allowing induction of a wrapper from a single hand-labeled instance per field per list. We then use the wrappers and data learned from the semi-supervised process to bootstrap an automatic (weakly supervised) wrapper induction process for additional lists in the same domain. In both induction scenarios, we automatically map labeled text to ontologically structured facts. Our implementation induces two kinds of wrappers, namely regular expressions and hidden Markov models. We evaluate our implementation in terms of annotation cost and extraction quality for lists in multiple types of historical documents.
* OntoSoar: There is a need to have an automated system that can read genealogy books or other historical texts and extract as many facts as possible from them. Embley and others have applied traditional information extraction techniques to this problem with a reasonable amount of success. In parallel much linguistic theory has been developed in the past decades, and Lonsdale and others have built computational embodiments of some of these theories. The goal of this thesis is to apply computational models of linguistic theory to the problem of extracting facts from genealogy books to find facts that traditional information extraction techniques could not find.
* GreenFIE-HD:
* WoK (including search & multilingual query processing)

Attachments: See papers and proposals about these projects at deg.byu.edu.