OntoSoar: Using Language to Find Genealogy Facts

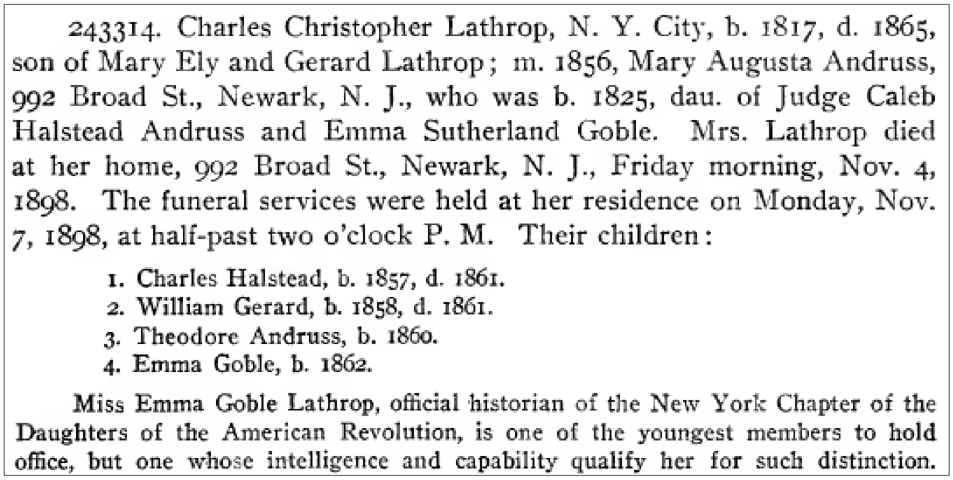
Peter Lindes, Deryle Lonsdale, David Embley
Brigham Young University

Abstract

OntoSoar is a system to extract information from genealogy texts and grow a conceptual model of the information. It is based on previous work embodied in systems called OntoES and LG-Soar. Here we describe the planned system, currently in early development, and give some preliminary and projected results.

Introduction

One potential large source of genealogical information is the thousands of historical books on family histories that have now been scanned and OCR'd. Figure 1 shows an example of part of a page from one of these books:



243314. Charles Christopher Lathrop, N. Y. City, b. 1817, d. 1865, son of Mary Ely and Gerard Lathrop; m. 1856, Mary Augusta Andruss, 992 Broad St., Newark, N. J., who was b. 1825, dau. of Judge Caleb Halstead Andruss and Emma Sutherland Goble. Mrs. Lathrop died at her home, 992 Broad St., Newark, N. J., Friday morning, Nov. 4, 1898. The funeral services were held at her residence on Monday, Nov. 7, 1898, at half-past two o'clock P. M. Their children:

1. Charles Halstead, b. 1857, d. 1861.
2. William Gerard, b. 1858, d. 1861.
3. Theodore Andruss, b. 1860.
4. Emma Goble, b. 1862.

Miss Emma Goble Lathrop, official historian of the New York Chapter of the Daughters of the American Revolution, is one of the youngest members to hold office, but one whose intelligence and capability qualify her for such distinction.

Figure 1: A Sample of Genealogy Text

This small sample contains dozens of facts about people, their names, their life events, and their family relationships. A collection of 50,000+ such books of several hundred pages each has many millions of facts, but it would take an enormous amount of work to extract all this information manually from the digital texts.

Previous work has been done on automating the extraction of this information. Embley et al (2011) discusses a system called OntoES that attacks the problem using a conceptual model and "extraction ontologies" that use regular expressions to find textual patterns that contain facts. Lonsdale et al (2007) describes a system based on LG-Soar that uses natural language processing techniques to find facts in the text.

Systems of this sort start with an explicit or implicit conceptual model of entities such as people, dates, and events and the relationships among these entities. The texts will often contain information not contemplated in this conceptual model; it would be desirable to make the model grow as we discover new entity classes or relations in the data. Cimiano (2006) and Wong (2012) survey the field of both ontology learning and information extraction from text.

The research described here extends the work of both Embley and Lonsdale by building a much more complete and robust version of Lonsdale's LG-Soar system as adapted for genealogy texts, and integrating that with Embley's OntoES system. The resulting system should be able to both extract many more facts and grow the ontology by learning new classes and relations. Figure 2 shows a block diagram of the resulting system:

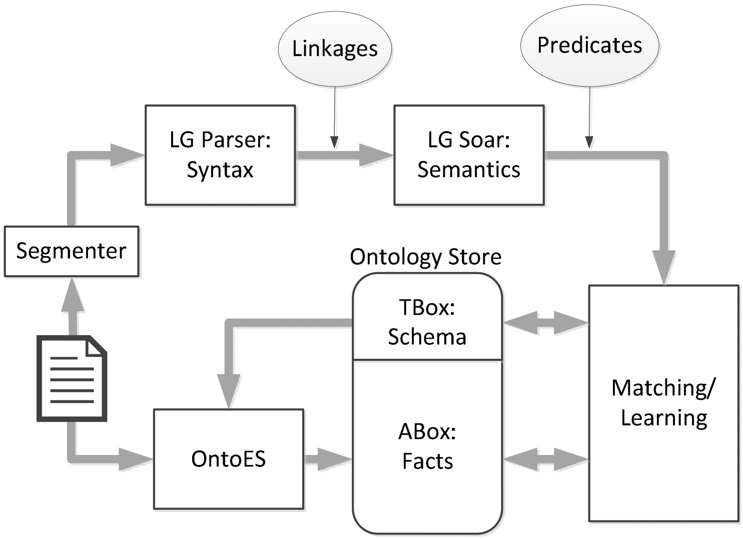


Figure 2: The OntoSoar System

The natural language processing part of the system segments the text into sentences and sentences fragments, parses these to produce syntax graphs called linkages, performs semantic analysis to derive predicates, and then integrates the results with the growing web of knowledge OntoES is building. At this writing the work is in the early stages, so we will show preliminary results and our plans for how we expect the finished system to work.

Syntactic Analysis

A brief examination of the text in Figure 1 shows that it is very non-standard English, with many lexical and grammatical abbreviations. Finding the information we want requires a solid syntactic analysis which connects words to build noun and prepositional phrases and shows the connections between verbs and their arguments. The Link Grammar Parser developed at CMU provides a robust English parser whose grammar is easily modified to adapt to the idiosyncrasies of genealogical text.

Figure 3 shows two examples sentence fragments and the linkages the LG parser produces. Since we are still in the process of adapting the grammar, these examples include some hand manipulation of the text to get the desired results.

Charles Christopher Lathrop, b. 1817, d. 1865,

```
    +---------------------------------Xp----------------------------------+
    |                            +------------Ss-----------+              |
    +-------------Wd-------------+--MX*p-+------Xc------+  |              |
    |        +----G----+----G----+    +Xd+--MVp--+-IN+  |  +--MVp--+-IN+  |
    |        |         |         |    |  |       |   |  |  |       |   |  |
LEFT-WALL Charles Christopher Lathrop , b.v [.] in 1817 , d.v [.] in 1865 .
```

m. 1856, Mary Augusta Andruss,

```
                  +---------MX---------+
                  |  +--------Xd-------+
  +-Ss+--MVp--+-IN+  |   +--G--+---G---+-Xc-+
  |   |       |   |  |   |     |       |    |
*GP* m.v [.] in 1856 , Mary Augusta Andruss ,
```

Figure 3: Syntactic Analysis

The first example shows the compound noun *Charles Christopher Lathrop* being the subject of *died*, and modified with a phrase about when he was *born*. The abbreviations of the two verbs were properly recognized. The proposition *in* was added by hand in both examples. In the second, a generic pronoun labeled **GP** was added as the subject of *married*, making it possible to complete this sentence and giving a point to link up later with the proper antecedent. As we complete our adaptation of the parser, these hand additions should be able to be taken care of automatically.

Semantic Analysis

The semantic analysis will take the linkages produced by the parser and proceed in three phases. The first phase transforms a linkage into a set of simple predicates for the words found. Versions of this phase have been implemented previously in some form in LG-Soar by Lonsdale et al (2001) and Wintermute (2012). This project is further enhancing and adapting it to the needs of genealogy.

The second phase uses a lexicon of verb frames to recognize people and events. For example, the verb *born* represents an event whose subject is a person and that may have a date attached. Finally, a third phase recognizes family relationships and builds family groups related to the people and events discovered in the second phase. Figure 4 shows an example based on the two fragments shown in Figure 3 plus a third one about a child.

Phase 1	Phase 2	Phase 3

Charles Christopher Lathrop, b. 1817, d. 1865.
in(born,N4)
in(died,N6)
1857(N4)
Charles(N2)
Christopher(N2)
Lathrop(N2)
born(N2)
1865(N6)
died(N2)
	NAME(N13,“Charles Christopher Lathrop”)
	DATE(D43,“1817”)
	DATE(D44,“1865”)
	isPerson(P22)
	hasName(P22,N13)
	isEvent(E7,“birth”)
	hasSubject(E7,P22)
	happenedIn(E7,D43)
	isEvent(E8,“death”)
	hasSubject(E8,P22)
	happenedIn(E8,D44)	

m. 1856, Mary Augusta Andruss.
in(married,N16)
1856(N16)
GNP(N15)
Mary(N18)
Augusta(N18)
Andruss(N18)
married(N15,N18)
	NAME(N21,“Mary Augusta Andruss”)
	DATE(D45,“1856”)
	isPerson(P23)
	hasName(P23,N21)
	isEvent(E9,“marriage”)
	hasSubject(E9,P22)
	hasObject(E9,P23)
	happenedIn(E9,D45)
		FAMILY(F27)
		createdBy(F27,E9)
		memberOf(P22,F27)
		memberOf(P23,F27)
		husbandIn(P22,F27)
		wifeIn(P23,F27)

Charles Halstead, b. 1857, d. 1861.
in(born,N25)
in(died,N26)
1857(N25)
Charles(N23)
Halstead(N23)
born(N25)
1861(N26)
died(N23)
	NAME(N29,“Charles Halstead”)
	DATE(D46,“1857”)
	DATE(D47,“1861”)
	isPerson(P24)
	hasName(P24,N29)
	isEvent(E10,“birth”)
	hasSubject(E10,P24)
	happenedIn(E10,D46)
	isEvent(E11,“death”)
	hasSubject(E11,P24)
	happenedIn(E11,D47)
		memberOf(P24,F27)
		childIn(P24,F27)
		NAME(N30,“Charles Halstead Lathrop”)
		hasName(P24,N30)

Figure 4: Semantic Analysis

Integration and Learning

Figure 5 shows a subset of the facts discovered by OntoES in our example paragraph. Comparing this with the predicates in Figure 4, we see that by finding matching lexical items we can match up osmx411, osmx415, and osmx478 with P22, P23, and P24, respectively. The corresponding birth and death dates can also be matched, allowing us to integrate the information from the two sources.

PersonName	Person	BirthYear	DeathYear
Charles Christopher Lathrop	osmx411	1817	1865
Mary Augusta Andruss	osmx415		
Charles Halstead	osmx478	1857	1861

Figure 5: OntoES Facts

Once this has been done, we should be able to find additional kinds of information that our original conceptual model does not represent. For example, the last part of the example in Figure 1 is free English text describing the professional history of the youngest child or our family. It should be possible to match up the names lexically, and then use our language model to determine that this is about her employment with a certain organization. Thus when we insert these facts into our web of knowledge we can grow the conceptual model to include a class of organizations and a relation that says they can employ people.

Evaluation

One reason for integrating the linguistic analysis of LG-Soar with the OntoES system is that OntoES also includes a number of associated tools. Among these tools is a web-based annotator that allows a person to mark up a page of text with the information on the names, dates, events, and family relationships found there. We intend to estimate the accuracy of our system by comparing its results with human annotation of a randomly selected sample of pages.

Conclusions

The OntoSoar system will greatly enhance the capability of the LG Parser and the LG-Soar system as a whole to build a powerful tool for analyzing genealogical texts. Through syntactic and semantic analysis and further inferencing based on the semantics we can discover information about people, their life events, and their family relationships. All this information can then be integrated with the conceptual model and facts in the OntoES system, making a full web of knowledge that will be available to the web-based tools in OntoES.

Our experience so far is that this new tool should enable us to find many more facts with high accuracy and also grow the conceptual model with new classes and relations.

References

Cimiano, Philipp (2006). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer, New York.

Embley, David W., Steven W. Liddle, and Deryle W. Lonsdale, (2011). “Conceptual Modeling Foundations for a Web of Knowledge”, in *Handbook of Conceptual Modeling*, Chapter 15.

Lonsdale, Deryle, Merrill Hutchison, Tim Richards, and William Tyson (2001). *An NLP system for extracting and representing knowledge from abbreviated text*. The BYU NL-Soar Research Group.

Wintermute, Sam (2012). “Leveraging Cognitive Context for Language Processing in Soar” in *The 32nd Soar Workshop*, June 2012, University of Michigan.

Wong, Wilson, Wei Liu, and Mohammed Bennamoun (2012). "Ontology Learning from Text" in *ACM Computing Surveys*, Vol. 44, No. 4, Article 20.