Near-Perfect Relation Extraction from Family Books

George Nagy(✉)

Rensselaer Polytechnic Institute, Troy, NY 12180, USA
nagy@ecse.rpi.edu

Abstract. Precise sequence constraints are proposed to accelerate information extraction from a class of "semi-structured" documents that includes hundreds of thousands of digitized genealogical records. While Named Entity Recognition (NER) and Named Relation Recognition (NRR) on free-running text lack universally applicable solutions, under these constraints generalized template-matching can accomplish both. Interactive information extraction is demonstrated on three digitized books. The book-text tokens are first labeled according to their role (e.g. Head, Spouse, or Birthdate), then pairs of labeled entities are combined into labeled relations (e.g. <Head [Spouse]>, or <Spouse [Birthdate]>). Accurate NRR is ensured by high-precision (>99%) NER. On semi-structured text the proposed NRR algorithm produces only valid relations from correctly labeled entities. About three hours of user interaction and a few minutes of laptop run time suffice to produce database- or ontology-ready input from a new book.

Keywords: Information extraction · Text analysis · Language models

1 Introduction

In response to the need for less laborious recovery of genealogical facts from printed family records, we present a model of semi-structured text that leads to simple and accurate information extraction by generalized template matching. Our program processes Unicode text with no formatting other than line breaks. The user needs to specify only a few exemplary templates and a list of the relations to be extracted.

Semi-structured family books typically contain interspersed *sign phrases* (like *born, died, spouse, son of*, or *dau. of*) and *value phrases* (*Henrietta Mills Hyde* or *1828*). A value may be located several lines away from its sign and span a variable configuration of tokens. The *GreenEx* collection of python modules scours the text for data needed to populate genealogical databases and ontologies.

The first pass over the text, *Named Entity Recognition* (NER), attaches each label (HEAD, SPOUSE, BIRTHDATE…) to its value. The second pass, *Named Relation Recognition* (NRR), aka *Semantic Relation Extraction,* links the labeled values into binary relations like <HEAD[CHILD]: Henry Hyde [Henrietta Mills Hyde]> or, from a subsequent record of Henry's daughter Henrietta's family, <SPOUSE [MARRIAGE-DATE]: Hyde Charles Smith Shelton, [1848]>. We tested the method on the 18th century

J. Lladós et al. (Eds.): ICDAR 2021, LNCS 12823, pp. 477–491, 2021.
https://doi.org/10.1007/978-3-030-86334-0_31

Kilbarchan parish register [1], the early 20th century Miller funeral home records [2], and The Ely Ancestry that spans three centuries [3] (illustrated in the Appendix).

The contributions we demonstrate are (1) a set of constraints imposed on text by the desired relations, and (2) a generalized template matching algorithm that extracts the specified relations from any text subject to the constraints. The experimental results confirm that the permissive semi-structure constraints obeyed by three diverse family books suffice for fast and accurate relationship extraction.

The next section is a review of relevant prior work. Section 3 defines the proposed semi-structure constraints. The following sections present Sect. 4. entity labeling, Sect. 5. relation extraction, Sect. 6. experimental results and Sect. 7. conclusions.

2 Prior Work

We discuss here only properties of relations of interest in information extraction. Among the pioneering achievements before 2000 were The Acquisition of Hyponyms [4], the NYU Proteus System (later extended to news, scientific papers and patents) [5], and Snowball for finding patterns in plain text [6].

A popular survey of relation extraction up to 2006 is Bach's and Badaskar's [7]. Their taxonomy, based on the amount of required human interaction and of relevant data, has stood the test of time. Zettlemoyer offers a lively introduction to both NER and NRR [8]. Dependency tree methods derive distance measures from grammatical relations between tokens [9]. The value of extending seed lists with unlabeled data is demonstrated in [10]. Joint NER-NRR was initiated in 2006 [11]. *Distant supervision* combines the advantages of supervised and unsupervised approaches by exploiting linguistic resources only indirectly related to the searched text [12, 13]. It is the focus of an authoritative CSUR review [14]. Approaches based on neural networks came along a little later [15]. Chen and Gu review NRR research up to 2019, catalog the shortcomings of existing methods, and compare their probabilistic joint NER-NRR to other methods on several biomedical benchmarks [16]. Many papers address only binary relations because in NER useful n-ary ($n > 2$) relations can often be factorized into binary relations [17]. *NLP-Progress*, a website that tracks developments in natural language processing, lists test data and competitions by language, model, and application [18].

A category-based language model is compared with a probabilistic finite-state machine model for labeling family roles in handwritten 17th Century Catalan marriage records in [19]. With a large fraction (6/7) of the 173-pages used for training, and seven-fold cross-validation, both methods yielded 70–80% Precision and Recall. Information extraction was also the topic of a 2017 ICDAR competition. Using neural networks and conditional random fields, the winning team (from Harbin Institute of Technology) achieved a remarkable character error rate of ~8% on the same database [20], but 100 of the 125 pages had to be manually labeled for training and validation. Combining NER/NR with HWR or OCR appears to be unique to the DAR community.

Our example-based approach is similar in spirit to end-user-provided training examples for scanned business documents [21]. Literals and semantic tags were anticipated in [22]. The effects of OCR errors on information retrieval were discussed in [23]. Recent shifts in the very nature of documents were reviewed in [24]. The popular Stanford

Named Entity Recognizer [25] failed on our books because it depends on probabilistic sentence analysis, but for tokenization we use the Natural Language Tool Kit (NLTK) that it spawned. Preceding the rapid rise of deep learning methods, rule-based extraction like ours was favored over machine learning in [26]. It remains to be seen whether a machine learning approach applied to semi-structured text can match generalized template matching with respect to user time, minimal training data, and accuracy.

The work closest to ours is that of the BYU and FamilySearch research team, which has access to 460,000 digitized publications of genealogical interest [27]. They describe in [28] a pipeline based on conceptual modeling for possible integration into FamilySearch's Family Tree. In [29] they present a thorough review of recent research on information extraction for genealogical purposes as well as some experiments on the same books as we used. The BYU team also proposed automated discovery of errors (like inconsistent dates) in sources of data [30]. They reported recovering family information from obituaries [31] and from lists abstracted from family books [32].

GreenEx was initiated to accelerate the construction of character-level REGEX templates at BYU [33]. The first versions of GreenEx lacked floating extracts, format variants and auto-suggestion routines and failed to exceed a Figure-of-Merit of 0.95 [34, 35]. (*Green* is for pattern recognition programs that never waste a user action, from [36]). Generalizing template matching yielded much better (~99%) entity recognition with fewer templates [37]. The development of accurate entity extraction on semi-structured genealogical records laid the foundation for relation extraction. We found no formal examination of semi-structure in the literature, and no comparable experiments.

3 Structure Constraints

Informally, *semi-structured documents* are lists of quasi-repetitive records where some tokens can be designated as signs or values of the information to be extracted. The, *sign* (a semiotic term) is also called *marker, query*, or *search phrase*, and *value* is *extract* or *target*. The subject of a record is called *Head*. Each f*amily record* is a sequence of phrases (*factoids*) about the Head's family. Figure 1 includes two family records (in lines 9–16 and 17–23) that we will use as a running example. Each record begins with a sign for the Head (here a six-digit record identifier) and extends to the first token of the next Head sign. In other books, the sign for Head is the name itself at the beginning of a line.

We define a binary relation *R* of *type* Label_1 [Label_2] as a pair of values (v_1 [v2]). An *instance* of a relation, from lines 10 and 11 of Fig. 1, is:

$R_{\text{SPOUSE [BIRTHDATE]}}$ = <SPOUSE [BIRTHDATE]; *Charles Smith Shelton* [*1819*]>

GreenEx reports every instance of $R_{\text{SPOUSE [BIRTHDATE]}}$.

We need some notation for the signs and values of *R* in semi-structured text. A sequence number (*SeqNo*) *N,* ranging from 1 to the number of word tokens in the book, is assigned to each word token. Let *s1, s2, v1*, and *v2* denote the signs and values designated to extract instances of relation *R*. Let *N(s1)* denote the SeqNo of the first token of *s1*, and *N(s2)* that of *s2*. Let *N0* be the SeqNo of the first token of the family record that contains *v1*, and *N1* the SeqNo of this record's last token. Figure 2 shows plausible signs and values for the genealogical factoids of a record in Fig. 1.

1. THE ELY ANCESTRY. 423
2. SEVENTH GENERATION.
3. b. 1826, d. 1857, son of Daniel Havens and Desire Holmes; m. 3rd,
4. 1862, Herbert Post, Marion, Ala., who was b. 1827, son of Truman Post
5. and Betsy Atwater. Their children:
6. 1. Robert Alexander, b. 1846; m. 1869, Katherine Pierce Parker.
7. 2. Julia Hyde, b. 1855.
8. 3. Etta Hyde, h. 1856, d. 1857.
9. 243357. Henrietta Mills Hyde (127 St. James Place, Brooklyn, N.
10. Y.), b. 1826, dau. of Julia Ely and Zabdial Hyde; m. 1848, Charles
11. Smith Shelton, Madura, India (missionary), b. 1819, d. 1879, son of
12. George Shelton and Betsy Wooster. Their children :
13. 1. Fanny Arabella, b. 1850; m. 1874, Arthur Harry Bissell.
14. 2. Julia Elizabeth, b. 1851 ; m. 1878, Chas. J. Van Tassel.
15. 3. Charles Henry, b. 1854, M.D., 288 Fourth St., Jersey City, N. responds.
16. 4. Henry Hyde, b. 1858.
17. 243358. Aurelia Carrington Hyde (127 St. James Place, Brooklyn,
18. N. Y), b. 1828, dau. of Julia Ely and Zabdial Hyde; m. 1848, Edward
19. Chauncey Halsey (Directory), 165 Warren St., Brooklyn, N. Y., who
20. was b. 1825. Their children :
21. 1. Eleanor Shelton, b. 1850.
22. 2. Adeline Sanford, b. 1852.
23. 3. Edward Carrington, b. 1854.
24. 243359. Zabdiel Sterling Hyde (care E. Goddard & Sons, 1020

Fig. 1. Part of a page of OCR'd text from the Ely Ancestry (line numbers and highlight for record # 243357 added)

243357. *Henrietta Mills Hyde* (127 St. James Place, Brooklyn, N. Y.), **b.** *1826*, **dau. of** *Julia Ely* and *Zabdial Hyde*; **m.** *1848*, *Charles Smith Shelton*, Madura, India (missionary), **b.** *1819*, **d.** *1879*, **son of** *George Shelton* **and** *Betsy Wooster*. Their **children** :
1. *Fanny Arabella*, **b.** *1850*; **m.** *1874*, *Arthur Harry Bissell*.
2. *Julia Elizabeth*, **b.** *1851* ; **m**. *1878*, *Chas. J. Van Tassel*.
3. Charles Henry, **b.** *1854*, M.D., 288 Fourth St., Jersey City, N. J., responds.
4. *Henry Hyde*, **b.** *1858*.

Fig. 2. Family record for Henrietta Mills Hyde, from the page shown in Fig. 1. Potential signs and values are colored green (bold) and red (italicized) respectively. (Color figure online)

Then the structure constraints on the text that suffice for a relation to be extractable are:

1. $N0 \leq \min(N(s1), N(s2), N(v1), N(v2)$, and $\max(N(s1), N(s2), N(v1), N(v2) \leq N1$ *(every sign and value participating in a relation must be in the same family record)*
2. $N(s1) \leq N(v1)$ and $N(s2) \leq N(v2)$ *(a value cannot precede its sign)*
3. $N(s1) \leq N(v1) \leq N(s2) < N(v2)$ *(signs and values alternate, except for collocations indicated by equality for "≤", where the sign and value share some or all tokens)*

4. If a relation *R1* associates *v1* and *v2*, then there cannot be a relation *R2* of the same type as *R1* between *v1* and *v2* (*relations of the same type cannot be nested*)

The following examples would violate these constraints:

Constraint #1: 243357. Henrietta Mills Hyde b. 243357. 1826 Abel
Constraint #2: 243357. Henrietta Mills Hyde 1826 b. 243357.
Constraint #3: 43357. b. Henrietta Mills Hyde 1826 243357.
Constraint #4: Henrietta Mills Hyde Fanny Arabella b. 1850 **b. 1826**

Constraint #1 implies that extracting inter-record relations requires further processing. #2 can be obviated with a reverse second pass. According to #3, either *m. 1848, Charles Smith Shelton* or *m. Charles Smith Shelton 1848* would be acceptable, with *m* serving in either case as the sign for both the *marriage date* 1848 and the *spouse* Charles Smith Shelton. This constraint occasionally requires some ingenuity in formulating the appropriate template (e.g. for twins with a single birth date). We have never seen a violation of #4.

The family record of Henrietta Mills Hyde (Fig. 2) is semi-structured with respect to every relation listed in Table 3 of Sect. 6. Semi-structure is a substitute for sentence structure to aid human comprehension.

4 First Level Template Matching

Generalized template matching achieves high precision entity recognition with few templates by (1) application-oriented (but not document-specific) *word tagging* based on alphanumeric format, (2) substituting common alternative noun and date configurations (*format variants*) for the ones specified in the template, and (3) extending the search for the value corresponding to a sign (*floating templates*) [37].

Recall is further improved by *auto-suggestion routines* that scan the book for tokens that should have been labeled but were not. Common causes of unlabeled tokens are unusual word configurations (like *John, in adultery 1675* instead of the expected *John, born 1675*), typesetting errors (often in punctuation), and OCR errors (*I* instead of *1*) that affect tagging. When the inconsistent text segments are displayed on a clickable form, the user can add a template that will correct the current error as well as similar errors elsewhere in the text [37].

Adding half-a-dozen templates (for each book) based on the suggestion routines raised recall by about a half percent with insignificant change in precision. Half percent is not negligible at Recall >98%. However, the effectiveness of adding templates gradually decreases: eventually each new template will correct only one or two errors.

When all recognizable tokens have been labeled, GreenEx assembles consecutive same-label tokens and their locations into *extract groups*. Then the extract groups of each family, bracketed by HEAD labels, are collected into *labeled family records*. The book-length list of labeled family records (e.g. Fig. 3) is the input to relation extraction.

…HEAD:,578,10,9,3,Henrietta,Mills,Hyde,B_DATE:,578,11,9,1,1826,PARENT1:,
578,11,23,2,Julia,Ely,PARENT2:,578,11,37,2,Zabdal,Hyde,M_DATE:,578,11,54,1,
1848,SPOUSE:,578,11,60,3,Charles,Smith,Shelton,B_DATE:,578,12,47,1,1819,
D_DATE:,578,12,56,1,1879,PARENT1:,578,13,1,2,George,Shelton,PARENT2:,
578,13,20,2,Betsy,Wooster,CHILD:,578,14,4,2,Fany,Arabella,B_DATE:,578,14,23,1,
1850,M_DATE:,578,14,32,1,1874,SPOUSE:,578,14,38,3,Arthur,Harry,Bissell,
CHILD:,578,15,4,2,Julia,Elizabeth,B_DATE:,578,15,24,1,1851,M_DATE:,578,15,34
,1,1878,SPOUSE:,578,15,40,5,Chas,.,J.,Van,Tsel,CHILD:,578,16,4,2,Charles,Henry,
B_DATE:,578,16,22,1,1854,CHILD:,578,17,4,2Henry,Hyde,B_DATE:,578,17,19,1,
1858,HEAD:, …

Fig. 3. Labeled family record of Henrietta Mills Hyde, including book coordinates: page, line, character and length (number of tokens)

5 Relation Extraction

Named Relation Recognition has been intensively studied for thirty years without finding a universal solution. The complexity of natural language requires complex language models, many training examples, or external resources for avoiding misses and errors. Our main point is that the task is much easier for text semi-structured with respect to the desired relations because every relation to be extracted is fully defined by the *labels* of the participating entities. Figure 4 shows pseudo-code for binary relation extraction from labeled family records.

In our notation, A[B] stands for a unique tuple within a family. Therefore relations like HEAD[CHILD] and CHILD[BIRTHDATE] must be understood as HEAD [FIRST-CHILD] or HEAD[SECOND-CHILD] and FIRST-CHILD[BIRTHDATE] or SECOND-CHILD[BIRTHDATE]. (N-ary relations are also restricted to *unique* tuples, and can therefore always be decomposed into dyads. Some authors exclude such tuples from the definition of n-ary ($n > 2$) relations.)

GreenEx extracts all the specified relations from one labeled family record at a time. In contrast to the NER pass, there is no need for tagging, format variants, provisions for line-ends and page breaks, or interactive template construction. The signs and values can only be the algorithmically assigned labels that define the relations. For a desired <Label_X [Label_Y]> relation, the program just loops over the label groups in each family to locate a Label_X group and the next Label_Y group. The values of these label groups constitute the sought relation. Therefore errors in relations can occur only when one of the constituent tag phrases was mislabeled or unlabeled in the NER pass.

Constraint #1 is satisfied by limiting the search for a value to the current family record. Restarting the search at the current sign satisfies *Constraint* #2. Halting the search before the next identical sign satisfies both #3 *and* #4. The stopping rules convert potential Precision errors in relation extraction (due to labels missed in the NER pass) to Recall errors. Therefore if the labels are correct, and the labeled text satisfies the constraints imposed by the specified relations, then only valid relations are extracted.

Twenty-six types of binary relations that can be extracted from our family books, under the semi-structure constraints on the current labels, are listed in Table 3. Shown below are some relations extracted from our running example. (We envy mathematicians,

whose notation for a binary relation is typically (*a,b*). In the first relation below, *a* is *Henrietta,Mills,Hyde* and *b* is *Charles,Smith,Shelton*). The following examples of extracted relations show relation type, book locations (page, line, token) of the values, and the values themselves. The book locations are *attributes* of the relations.

Two examples of an instance of a user-specified binary relation:

<HEAD [SPOUSE] 57810 9, 578 11 60 Henrietta,Mills,Hyde [Charles,Smith,Shelton]>

<CHILD [SPOUSE] 578 15 4, 578 15 40 Julia,Elizabeth [Chas,.,J.,Van Tassel]>

```
Function Grelex (FamilyRecords, DesiredTuples)
   Input:   FamilyRecords, DesiredTuples
   Output: ExtractedRelations

Convert FamilyRecords to Families          % Family Records is a book-length list
                                             of labeled groups of tokens
For Family in Families:      % Family is a list of Groups, each [Label, Page, Line,
      Offset, Value], in a single Family; Families is a book-length list of Family(s)
   For XGroup in Family                    % restrict search to this Family
      Excerpt XLabel, XValue, XID from XGroup
                                          % XID_ is Page, Line, Offset of XGroup's value
      For Tuple in DesiredTuples:   % for every specified relation
         Excerpt LeftLabel, RightLabel from Tuple  % set the two search arguments
         StopGroup ← Stopper(LeftLabel, RightLabel)
                                          % StopGroup is a context-dependent label
         If XLabel =LeftLabel     % if an entity label matches left label of this tuple
            For YGroup in Family from XGroup to StopGroup :
                                % search forward to find a match for the right label
               If YLabel = RightLabel  % if a match for right label is found
                  Excerpt YLabel, YValue, YID from YGroup
                  Relation ← [[XLabel, YLabel, XID YID],[XValue],[YValue]]
                                   % extract this instance of the desired relation
                  Append Relation to ExtractedRelations
               End If YLabel = RightLabel
            End  For YGroup
         End  If XLabel =LeftLabel
      End  For Tuple
   End  For XGroup
End  For Family
Return(ExtractedRelations)   % ExtractedRelations is a list of attributed relations
      of the form Label_1[Label_2], , ID_1, ID_2, Value_1[ Value_2]
```

Fig. 4. Simplified pseudocode for extracting binary relations from family records. Code for stopping rules with provisions for multiples (e.g. spouses, children, twins) and alternatives (birth-date, christening date) not shown.

The derivation of a decomposable n-ary relation from its constituent binary relations is straightforward. A recursive GreenEx routine factors each n-ary relation into binary relations, e.g. A[B[C]] into A[B] and B[C], or A[B,C] into A[B] and A[C]. The program then fills the slots of the n-ary relation with the elements of the already extracted binary relations. Many important applications (e.g. populating relational databases and Resource Description Framework RDF triples) require only binary relations. The experiment below is confined to attributed binary relations. Two examples of an instance of a user-specified n-ary relation:

<HEAD [CHILD, [B_DATE]] 57810 9, 578 15 40, 578 15 24; Henrietta,Mills,Hyde [Julia,Elizabeth [1803]]>

<HEAD, [SPOUSE [[PARENT1], [PARENT2]]] 400 13 08, 400 14 32, 400 15 27, 400 15 48; Henrietta,Mills,Hyde [Charles,Smith,Shelton [[George,Shelton], [Betsy,Wooster]]]>

6 Experimental Results

Table 1 shows the results of processing the three books, and the accuracy on the manually labeled test data. The tokens labeled "NONE", like addresses, occupations, military ranks, officiating clergy, and the names of informants, were excluded from the labeled family records. They were included as an additional class in the precision and error calculations. In two of the books Precision is 99.9%, and in the third it is 99.7%. We note, however, that first-stage labeling failure of a single shared value could cause missing several relations. Table 2 summarizes the results of relation extraction. Table 3 displays the number of relations extracted from each book by relation type.

We did not find any instance of the 26 relations listed that failed to obey the constraints. All missed relations in the ground-truthed pages were due to OCR errors or misprints (but some were caught by the suggestion routines). Extracting almost all of

Table 1. Data characteristics and accuracy

	Kilbarchan	Miller	Ely
Pages processed	139	389	301
Specified labels	8	10	8
Tokens assigned specified labels	39203	91633	39440
Tokens labeled NONE	34077	131462	101485
Test Set with ground truth			
Pages	6	6	6
Tokens (including "NONE")	3126	3842	3423
Precision	0.999	0.999	0.997
Recall	0.981	0.991	0.992
F-measure	**0.990**	**0.996**	**0.994**

the desired information from a new book takes less than three hours of interactive template construction and only a few minutes runtime on a 2.4-GHz Dell Optiplex 7010 with 8-GB RAM running Python 3.6 with IDLE under Windows 10.

Table 2. Summary of relation extraction

	Kilbachan	Miller	Ely
Types to extract	26	26	26
Types found	15	17	16
Families	2615	4186	1219
Extracted Groups	14152	38146	20781
Extracted Relations	**11487**	**25633**	**18780**

Table 3. Number of extracted relations of each type

Relation	Kilbarchan	Miller	Ely
HEAD[CHILD]	4076	837	3448
HEAD[TWINS]	47	0	0
HEAD[SPOUSE]	2102	1730	1306
HEAD[B_DATE]	6	2849	1128
HEAD[M_DATE]	963	0	1020
HEAD[D_DATE]	0	4119	374
HEAD[BU_DATE]	0	3524	0
HEAD[B_PLACE]	0	2798	0
HEAD[M_PLACE]	134	0	0
HEAD[BU_PLACE]	0	4053	0
HEAD[PARENT1]	0	2824	1203
HEAD[PARENT2]	0	2726	1210
SPOUSE[B_DATE]	3	9	1003
SPOUSE[M_DATE]	12	0	136
SPOUSE[D_DATE]	0	8	360
SPOUSE[B_PLACE]	0	10	0
SPOUSE[M_PLACE]	12	0	0
SPOUSE[PARENT1]	0	9	1015
SPOUSE[PARENT2]	0	7	1028
CHILD[B_DATE]	1143	43	3282
CHILD[C_DATE]	2873	0	0
CHILD[D_DATE]	0	82	831
CHILD[SPOUSE]	34	5	819
CHILD[M_DATE]	38	0	617
TWINS[B_DATE]	6	0	0
TWINS[C_DATE]	38	0	0
TOTAL	**11487**	**25633**	**18780**

7 Conclusion

What we learned from the experiments is the unexpected simplification of Named Entity Recognition and Named Relation Extraction enabled by appropriate characterization of semi-structured text. The constraints listed in Sect. 3 proved just tight enough to allow generalized template recognition to yield much higher precision and recall on both tasks than reported by others on free-flowing test data, and loose enough to fit the three diverse books recommended to us for testing. These books differed not only from each other because of purpose and date, but also internally because they were compiled over several lifetimes by many authors.

Particularly gratifying was the discovery that template matching enables error-free binary relation extraction on correctly labeled semi-structured text. This is assured because template matching just maps the extracted entity labels into a highly redundant list of relations without using or introducing any external information. No such claim can be made for free-flowing text.

Template matching is linear in the length of the input, so checking text compliance with the rules directly would be only slightly faster than running GreenEx. Template construction is necessarily book-specific. For example, in Miller, *m* indicates *mother*, but in Kilbarchan and Aly it points to *married.* Fortunately, we were able to show in earlier papers that customizing the system to each book (with appropriate computer help and a user-friendly interface) requires surprisingly little human interaction. We expect, but have not proved, that the skill level required is within reach of most current users of genealogical software.

No machine learning was tried or compared. With all F-scores $\geq$99%, what could any comparison on the same data prove? Avoiding the laborious preparation of training data is the main point of the proposed approach. No comparison with statistical classifiers, including deep learning, can contest that. From the perspective of genealogists it seems more urgent and useful to determine what fraction of the plethora of family books obeys the postulated semi-structure constraints.

The most serious potential error at the NER stage is a missed Head. This could assign a Child, a Parent, or a Spouse to the preceding Head, and consequently yield some incorrect relations. Although there were some OCR errors on Heads in our test set, they were all caught by the auto-suggestion routines.

The results could be filtered by genealogy-specific checks to detect missing names (every person with some attribute must have one), missing birthdates (in Ely, every Head has one), more than two parents, inconsistent birth, marriage and death dates, and other definite or suspect genealogical inconsistencies. We don't, however, have any dependable method for automatic correction of detected errors. We expect the most significant advance to come from larger scale projects that combine results from multiple genealogical sources covering the same community.

Only part of the simple tagging routine in GreenEx is specific to English family books. Therefore the proposed method could perhaps be extended to other semi-structured books of historical interest like city directories and product or merchandise catalogs, and to other languages and scripts. As it stands, the only contribution claimed is a simple and effective method of named relation extraction from family books.

Acknowledgment. I am grateful to Emeritus BYU Professor David E. Embley and his colleagues for the digitized family books and for their sustained interest, advice and critique. The cogent suggestions of the three ICDAR reviewers prompted appropriate revisions.

Appendix I. Sample of Text from the Kilbarchan Parish Register

```
Parish of Kilbarchan.
Adame, Robert, par., and Issobell Adame, par. of Loch-
winnoch, in Pennell 1679 m- 2I Mar. '678
A daughter, 30 Mar. 1679.
Adam, William, par., and Elizabeth Alexander, par. of Paisley
m. Paisley, 15 May 1650
Adamson, Alexander, in Kilbarchan, and Mary Aitken p. 12 Feb. 1763
Mary, born 16 Oct. 1763.
David, born 1 May 1765.
Aird, William, and Margaret Aitken, in Auchincloigh
Margaret, 9 Feb. 1707.
Aitken (Akin), and Elspa Orr m. 18 Dec. 1693
Aitken, Allan, and Mary Aitken
Agnes, 10 May 1741.
Aiken, David, and Janet Stevenson m. 29 Sept. 1691
Aitkine, Thomas, and Geills Ore m. 21 Dec. 1661
W. Richard Allasone and Ninian Aitkine.
Aikine, James, and Jean Allason, in Ramferlie, 1696 in Kaimhill
m. 23 Jan. 1679
John, 28 Nov. 1679.
William, 28 Aug. 1681.
Isobel, 12 July 1691.
Thomas, 19 Jan. 1696.
Allan, 19 July 1698.
Elizabeth, 26 May 1701.
Aitken, James, in Sandholes, and Mary Henderson p. 10 July 1741
John, 26 Dec. 1742.
James, born 28 Sept. 1744.
Robert, born 12 May 1747.
Matthew, born 11 April 1749.
William, born 22 April 1756.
Aitken, James, and Janet Moodie
Elizabeth, 25 July 1742.
Aitken. James, in Abbey par. of Paisley, and Janet Lyle, par.
p. 7 June 1755
Aitken, James, in Kilbarchan, and Jane Lindsay
Jane, born 4 July 1755.
Margaret, born 9 Sept. 1757.
John, born 21 Oct. 1762.
James, born 26 Oct. 1764.
Janet, born 14 April 1768.
Aitken, James, in Lochermiln, and Janet Gardner 1760 in Barbush
Mary, born 15 May 1758.
Christian, born 1 May 1760.
Janet, born 19 Jan. 1764.
Robert, born 28 May 1766.
Jane, born 28 Aug. 1768.
```

Aitken, James, par., and Janet Houstoun, par. of Houstoun p. 11 Jan. 1772
Aitkine, John, par., and Janet Muire, par. of Paisley
m. Paisley, 22 Oct. 1650
Akine, John, 1655 in Todhils
William, 15 Oct. 1652.
James and William, 9 April 1654.
Jonet, 1 July 1655.
Margaret, 1 May 1659.

Appendix II. Sample of Text from the Miller Funeral Home Records

ABERNATHY, ELMER d 4 April 1924 252 Bellevernon Ave BD Greenville Cem 6 Apr 1924 b 5 Oct 1863 age 60-5-29 pd by ELLEN ABERNATHY
ACCETTE, FRANK d 16 Oct 1942 Friday 3:15p.m. Greenville Dke Co OH BD Oct 1942 Abbottsville Cem Dke Co OH b 20 April 1897 Montreal Canada age 45-5-26 f JOSEPH ACCETTE m AGNES QUEIRLLON waiter in restaurant sp & informant LENA ACCETTE 405 Central Ave physician Dr Mills religion Catholic War record: enlisted 21 Feb 1918 disch 19 Aug 1919 World War I Canadian Expeditionary Force Army 2nd Depot Batt C.O. Reg . Services Catholic Church clergy Father Gnau
ADAMS, ADAM DANIEL d 30 Aug 1931 Miami Valley Hosp Dayton OH BD Castine Cem 2 Sept 1931 b 18 May 1872 Preston Co WV age 59-3-12 f COLEMAN AD.AMS Barber Co WV m RACHAEL BOWMAN Barber Co WV married farmer
ADAMS, ANNA E. 2215 Rustic Road Dayton OH d 24 Aug 1942 Monday 2:45a.m. Dayton Montgomery Co OH BD 26 Aug 1942 Frankli n Cem OH b 7 Feb 1856 Franklin OH age 86-6-17 f DAVID ADAMS single housekeeper informant Mrs LOUIS MEYERS 2215 Rustic Road Dayton OH physician Dr Sacks clergy Rev Jones Dayton OH services Baptist Church in Franklin OH
ADKINS, HESTER d 6 Nov 1925 Weaver's Station BD Fort Jefferson Cem 8 Nov 1925 age 84-3-9 chg to RILEY ADKINS, pd by JAMES A. ADKINS
ADKINS, JAMES ALEXANDER d 2 Sept 1944 Wayne Hosp Greenville OH BD 4 Sept 1944 Fort Jefferson Cem Dke Co OH b 19 June 1872 Vandalia IL age 72-2-13 f RILEY ADKINS Dke Co OH m HESTER McCOOL retired rural mail carrier sp CORA ADKINS 65 years sisters Mrs MARY VIETS Dayton & Mrs CLATE RIEGLE Fort Jefferson
AIKEY, JACOB CLARENCE d 2 Oct 1937 N.W. of Pikeville 1~ mile BD Oakland Cem 5 Oct 1937 b 14 Dec 1855 Union Co PA age 81-9-18 f THOMAS AIKEY Maine m ALVINA KATHERMAN married retired farmer sp LYDIA
AIKEY, LYDIA ANN d 27 Aug 1925 7~ mile N.E. of Greenville BD Oakland Cem 30 Aug 1925 age 60-9-23 chg to JACOB AIKEY
ALBRIGHT, ADAM C. d 28 June 1920 Piqua OH hosp BD Abbottsville Cem 1 July 1920 age 72-7-27
ALBRIGHT, CARL ROLAND d 21 June 1917 VanBuren Twp BD Abbottsville Cem 23 June 1917 b VanBuren Twp age 11-2-10 f ALLEN ALBRIGHT m ANNA WEAVER
ALBRIGHT, CATHARINE d 10 May 1930 4 mile S.W. BD Greenville Mausoleum 13 May 1930 age 94-5-20 pd by DAYTON & CHAS ALBRIGHT
ALBRIGHT, ESTHER R. d 1 Jan 1946 113 Sherman St Dayton OH BD Abbottsville Cem Dke Co OH 3 Jan 1946 b 22 July 1863 Butler Co OH age 82-6-9 f THOMAS BENTON MORRIS Butler Co OH m ANGELINE HARROD Hamilton Co OH housekeeper widow sp WINFIELD S. ALBRIGHT 1 daughter Mrs HENRY RANCH 4 sons HENDER-SON of Greenville WILBUR of Greenville GEO of Dayton ELBERT of Dayton 12 grandchildren 2 brothers ARTHUR MORRIS Venice OH & SAM MORRIS Harrison OH 2 sisters Miss ELLA MORRIS Greenvi lle OH & Mrs ADA HARP Tulsa OK
1

Appendix III. Sample of Text from the Ely Ancestry

```
422 THE ELY ANCESTRY.
SEVENTH GENERATION.
243331- James Joseph Ernest Ely, son of Elisha Mills Ely and
Catherine Elizabeth Boode; m. Anna Horloff. Their children:
1. Alphonse.
2. August.
3. Alice.
4. Alfred.
243332. Alphonso Ethelbert Mills Ely, Palmyra, Mo., b. 1821, son
of Elisha Mills Ely and Catherine Elizabeth Boode; m. 1841, Drusilla
Pinkston, Palmyra, Mo., who was b. 1820, dau. of Peter Pinkston and
Abig-ail Davis. Their children :
1. Laura Ann Catherine, b. 1842.
2. Emma McLellan, b. 1850.
3. Alphonse Ethelbert Mills, b. 1852.
4. Mary Bailey, b. 1855.
5. Ophelia Goldburg, b. 1861.
243351. Elizabeth Plummer Hyde, b. 1814, d. 1855, dau. of Julia
Ely and Zabdial Hyde; m. 1834, Robert McClay Henning (243362X),
who was b. 1812, d. 1875, son of James Gordon Henning and Alicia
Courtney Spinner. Their children:
1. James Spencer, b. 1835.
2. Julia Ely, b. 1837; m. 1854, Robert Pearce (Mrs. Julia Ely Pearce, 58
St. John's PL, Brooklyn).
3. Edwin Courtney, b. 1838.
4. Henrietta Mills, b. 1841. (Her son is Dr. Chas. H. Shelton, 288 Fourth
St., Jersey City).
5. Elizabeth Alicia, b. 1843.
6. Robert McClay, b. 1847.
7. James Woodruff, b. 1850.
243353. Edwin Clark Hyde (13 Warren St., St. Louis), b. 1819,
son of Julia Ely and Zabdial Hyde ; m. 1844, Elizabeth Ann Peake (Gor-
don), who was b. 1816, dau. of Henry Peake and Isabella Herring
( Snyder) . Their children :
1. Henrietta Mills, b. 1845, d. 1848.
2. Susan Isabella, b. 1847.
3. Samuel Peake, b. 1850.
4. Annie Carroll, b. 1851, d. 1857.
5. Allen Withers, b. 185S, d. 1856.
243356. Julia Ely Hyde, Marion, Perry Co., Ala., b. 1824, 'dau. of
Julia Ely and Zabdial Hyde; m. 1844, Alexander Clark Bunker, who
(vas b. 1822, d. 1846, son of Thomas Bunker and Sally Raymond; m.
2nd, 1854, Washington Holmes Havens, St. Francisville, Mo., who was
```

References

1. Grant, F.J. (ed.): Index to the Register of Marriages and Baptisms in the PARISH OF KILBARCHAN, pp. 1649–1772. J. Skinner & Company, Ltd, Edinburgh (1912)
2. Miller Funeral Home Records, 1917–1950, Greenville, Ohio. Darke County Ohio Genealogical Society, Greenville (1990)
3. Vanderpoel, G.: The Ely Ancestry: Lineage of RICHARD ELY of Plymouth, England. The Calumet Press, New York (1902)

4. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France (1992)
5. Grishman, R., Sterling, J., Macleod, C.: Description of the Proteus system as used for MUC-3. In: Proceedings of the Third Message Understanding Conference, San Diego, CA, pp. 183–190 (1991)
6. Agichtein, E., Gravano, L.: Snowball: extracting relations from large plain-text collections. In: Fifth ACM Conference on Digital Libraries, San Antonio, TX (2000)
7. Bach, N., Badaskar, S.: A Review of Relation Extraction (2006). https://www.cs.cmu.edu/~nbach/papers/A-survey-on-Relation-Extraction.pdf
8. Zettlemoyer, L.: Relation Extraction (2013). https://docplayer.net/31229549-Relation-extraction-luke-zettlemoyer-cse-517-winter-2013.html
9. Culotta, A., Sorensen, J.: Dependency tree kernels for relation extraction. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, pp. 423–429 (2004)
10. Talukdar, P.P., Brants, T., Liberman, M., Pererira, F.: A context pattern induction method for named entity extraction. In: Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X), New York City, pp. 141–148 (2006)
11. Choi, Y., Brock, E., Cardie, C.: Joint extraction of entities and relations for opinion recognition. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pp. 431–439, July 2006
12. Zettlemoyer, L.: Advanced Relation Extraction (2013). https://cs.nyu.edu/courses/spring17/CSCI-GA.2590-001/DependencyPaths.pdf
13. Min, B., Grishman, R., Wan, L., Wang, C., Gondek, D.: Distant supervision for relation extraction with an incomplete knowledge base. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, pp. 777–782 (2013)
14. Smirnova, A., Cudré-Mauroux, P.: Relation extraction using distant supervision: a survey. ACM Comput. Surv. (2018). Article no. 106
15. Cai, R., Zhang, X., Wang, H.: Bidirectional recurrent convolutional neural network for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, pp. 756–765 (2016)
16. Chen, J., Gu, J.: Jointly extract entities and their relations from biomedical text. IEEE Access **7**, 162818–16227 (2019)
17. McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y., White, P.: Simple algorithms for complex relation extraction with applications to biomedical IE. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, MI, pp. 491–498 (2005)
18. NLP-Progress. http://nlpprogress.com/. Accessed 15 Mar 2020
19. Romero, V., Fornes, A., Vidal, E., Sanchez, J.A.: Using the MGGI methodology for category-based language modeling in handwritten marriage licenses books. In: Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (2016)
20. Fornes, A., et al.: ICDAR 2017 competition on information extraction in historical handwritten records. In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (2017)
21. Schuster, D., et al.: Intellix – end-user trained information extraction for document archiving. In: Proceedings of the International Conference on Document Analysis and Recognition, Washington (2013)
22. Sutherland, S.: Learning information extraction rules for semi-structured and free text. Mach. Learn. **34**, 232–272 (1999)

23. Taghve, K., Nartker, T.A., Borsack, J.: Information access in the presence of OCR errors. In: Proceedings of the ACM Hardcopy Document Processing Workshop, Washington, DC, pp. 1–8 (2004)
24. Nagy, G.: Disruptive developments in document recognition. Pattern Recogn. Lett. (2016). https://doi.org/10.1016/j.patrec.2015.11.024
25. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363–370 (2005)
26. Chiticariu, L., Li, Y., Reiss, F.R.: Rule-based Information Extraction is Dead! Long Live Rule-based Information Extraction Systems! Seattle, Washington, USA, pp. 827–832 (2013)
27. Family History Archives. https://www.familysearch.org/blog/en/family-history-books/. Accessed 21 Mar 2020
28. Embley, D.W., Liddle, S.W., Eastmond, S., Lonsdale, D.W., Woodfield, S.N.: Conceptual modeling in accelerating information ingest into family tree. In: Cabot, J., Gómez, C., Pastor, O., Sancho, M. (eds.) Conceptual Modeling Perspectives. Springer, Cham, pp. 69–84 (2017). https://doi.org/10.1007/978-3-319-67271-7_6
29. Embley, D.W., Liddle, S.W., Lonsdale, D.W., Woodfield, S.N.: Ontological document reading. An experience report. Int. J. Concept. Model. **13**(2), 133–181 (2018)
30. Woodfield, S.N., Seeger, S., Litster, S., Liddle, S.W., Grace, B., Embley, D.W.: Ontological deep data cleaning. In: Trujillo, J.C., et al. (eds.) ER 2018. LNCS, vol. 11157, pp. 100–108. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00847-5_9
31. Schone, P., Gehring, J.: Genealogical indexing of obituaries using automatic processes. In: Proceedings of the Family History Technical Workshop (FHTW 2016), Provo, Utah, USA (2016). https://fhtw.byu.edu/archive/2016
32. Packer, T.L., Embley, D.W.: Unsupervised training of HMM structure and parameters for OCRed list recognition and ontology population. In: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, Nancy, France, pp. 23–30 (2015)
33. Kim, T.: A green form-based information extraction system for historical documents. MA thesis, Brigham Young University, Provo, Uta (2017)
34. Embley, D.W., Nagy, G.: Green interaction for extracting family information from OCR'd books. In: Proceedings of the Document Analysis Systems Workshop (DAS 2018), Vienna (2018). https://doi.org/10.1109/DAS.2018.58
35. Embley, D.W., Nagy, G.: Extraction Rule Creation by Text Snippet Examples, Family History Technology Workshop, Provo, UT (2018)
36. Nagy, G.: Estimation, learning, and adaptation: systems that improve with use. In: Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, Hiroshima, Japan, pp. 1–10 (2012)
37. Nagy, G.: Green information extraction from family books. SN Comput. Sci. **1**, 23 (2020)