Automatic Metadata Extraction for Archival Description and Access

William Underwood
Georgia Tech Research Institute

Society of American Archivists
Research Forum
San Francisco
August 26, 2008
Research Motivation

• In responding to FOIA requests, Archivists need to be able to search collections of records with high precision and recall.
  • But at the time of responding to FOIA requests, archivists have not read all of the records, so cannot index the records and search on such attributes as person, organization and location names, topics, dates, author’s and addressee’s names and document types.

• Archivists cannot describe a collection until the collection has been manually read and reviewed.
  • With increasing volumes of electronic records, it may be decades or even centuries before new acquisitions are described.
Research Objectives

• Techniques for automatically recognizing document types and extracting metadata from electronic records that can be used for
  • indexing and searching collections of records by person, organization and location names, topics, dates, author’s and addressee’s names and document types, and for
  • automatically describing items, file units and record series.
Document Types: Examples in Presidential E-Records

Agenda
Bar Chart
Biography
Briefing Memo
Decision Memo
Correspondence
Diary
Executive Order
Information Memo
Job Application
List of Candidates for Federal Office
Mailing List
Memo
Minutes of Meeting
National Security Directive (NSD)

Newsletter
Nomination to Federal Office
Notes
Presidential Statement
Press Pool Report
Press Release
Referral Memo
Resume
Schedule
Signature Memo
Situation Report
Summary
Transcript of Speech
Telephone Call Recommendation
Transcript of News Conference
Document Types: Definitions

Documentary form is “the rules of representation used to convey a message, that is, the characteristics of a document which can be separated from the determination of the particular subjects, or places it concerns. Documentary form is both physical and intellectual.”

The intellectual form of a document is "the sum of a record's formal attributes that represent and communicate the elements of the action in which the record is involved and of its immediate context, both documentary and administrative."

The physical form of a document is “the overall appearance, configuration, or shape, derived from its material characteristics and independent of its intellectual content.”

[L. Duranti, Diplomatics: New Uses for an old Science]
Method for Recognizing Document Types and Extracting Date, Author, Addressee and Topic

1. Document Reader
2. English Tokenizer
3. Wordlist Lookup + *enhanced wordlists*
4. Sentence Splitter
5. Hepple POS Tagger + lexicon
6. Semantic Tagger + *Named Entity Rules*
7. *Intellectual Element Annotator* + *Intellectual Element Rules (DER)*
8. SUPPLE Parser/Interpreter + *Document Type Grammars augmented with Semantics*
9. *Extract Metadata*
Information Extraction: Wordlists

- Person_female_first.lst (8263)
- Person_male_first.lst (3704)
- Person_surname.lst (83,805)
- Location_city_us.lst (33,017)
- Location_us_county.lst (1,938)
- Location_us_state.lst (50)
- Location_foreign_city.lst (3802)
- Location_country.lst (458)
- Org_noun.lst (915)
- Org_ending.lst (238)
- Org_us_govt_dept_agency.lst (519)
Java Annotation Pattern Engine (JAPE) Rules

Rule: PersonMiddleInitial
Priority: 95
// Donald J. Atwood
// Mr. William H. Taft
{
    (TITLE)?
    (FIRSTNAME) | FIRSTNAMEAMBIG | LASTNAMEAMBIG)
    (NAME_INITIALS)
    (LASTNAME | LASTNAMEAMBIG | UPPER)|
    (PERSONENDING)?
}:person
--> :person.TempPerson = {kind = "personName", rule = "PersonMiddleInitial"

Rule: LocationCityCountry
// Sydney, Australia
// New York, United States
// Beijing, China
// This rule helps identify
// ambiguous foreign city names
Priority: 125
{
    ({Lookup.majorType == location,
      Lookup.minorType == city_foreign_ambig} |
    {Lookup.majorType == location,
      Lookup.minorType == city_foreign}
    }:locName
    ({Token.string == ","})|
    ({Lookup.majorType == location,
      Lookup.minorType == country})
}
--> :locName.TempLocation =
    {kind = "locName", rule = LocationCityCountry}
Examples of Intellectual Element Rules

for → MEMORANDUM FOR
from → FROM:
subj → SUBJECT:
para → <paragraph> </paragraph>
date → <Date> </Date> and not in a paragraph
attachment → Attachments
attachment → Attachment
Documentary Form: Intellectual Element Recognition

<document>

<chrondate>April 27, 1992</chrondate>

<for>MEMORANDUM FOR</for> <person>SAUL SKINNER</person>

<from>FROM: </from> <person>EME HOLIDAY</person>

<subject>SUBJECT: </subject> <topic>California Earthquake</topic>

<para>Attached is a situation report from FEMA on the northern California earthquake. No deaths have been reported and 45 people are known to have suffered injuries. In addition, there has been extensive property damage. While FEMA is awaiting a request from the State before initiating any recovery activities, a joint State/Federal preliminary damage assessment is likely to begin today.</para>

<para>Director Stickney has requested that we forward the situation report to you.</para>

<attachment>Attachments</attachment>

</document>
Document Types: Grammar for a Memorandum

MEMO → MEMOHEAD BODY
MEMO → MEMOHEAD BODY OPTIONAL
MEMOHEAD → DATE ADDRLINE SNDRLINE SUBJLINE
MEMOHEAD → DATE ADDRLINE THRULINE SNDRLINE SUBJLINE
ADDRLINE → FOR ENTITIES
SNDRLINE → FROM ENTITIES
SUBJLINE → SUBJ TOPIC
THRULINE → THRU ENTITY
BODY → PARAS
OPTIONAL → ATTACHMENT CCLIST BCCLIST
OPTIONAL → ATTACHMENT BCCLIST
OPTIONAL → ATTACHMENT CCLIST
OPTIONAL → ATTACHMENT
OPTIONAL → CCLIST BCCLIST
OPTIONAL → BCCLIST
OPTIONAL → CCLIST
CCLIST → CC ENTITIES
BCCLIST → BCC ENTITIES
PARAS → PARA PARAS
PARAS → PARA
ENTITIES → ENTITIES ENTITY
ENTITIES → ENTITY
ENTITY → PERSON JOBTITLE
ENTITY → JOBTITLE
ENTITY → PERSON
Grammar for Memorandum Augmented with Semantic Rules

%%% MEMO --> MEMOHEAD BODY
rule(memo(s_form:F, sem:D^E2^E1^[[document,D],
            [document_form,D, 'White House Memorandum'], [author,D,E2],
            SNDRList,[addressee,D,E1],ADDRList,[topic,D,TOPI]),
            [date,D,DATE]],
            [memohead(s_form:F, sem:E1^E2^[DATE,ADDRList,SNDRList,TOPI]),
             body(s_form:F)]).

%%% MEMOHEAD --> CHRONDATE ADDRLINE SNDRLINE SUBJLINE
rule(memohead(s_form:F, sem:E1^E2^[DATE,ADDRList,SNDRList,TOPI]),
     [chrondate(s_form:F, sem:DATE),
      addrline(s_form:F, sem:E1^ADDRList),
      sndrline(s_form:F, sem:E2^SNDRList),
      subjline(s_form:F, sem:TOPI)]).

%%% ADDRLINE --> FOR ENTITIES
rule(addrline(s_form:F, sem:ADDRList),
     [for(s_form:F), entities(s_form:F, sem:ADDRList)]).

%%% ENTITIES --> ENTITY
rule(entities(s_form:F, sem:E^SEM),
     [entity(s_form:F, sem:E^SEM)]).

%%% ENTITY --> PERSON
rule(entity(s_form:F, sem:E^[name,E,PERS]),
     [person(s_form:F, sem:PERS)]).
Parse Tree and Semantics of the Document

```
{best_parse=(memo
    (head (chrondate (sem_cat "April 27, 1992")))
    (addrline (for (sem_cat "MEMORANDUM FOR"))
       (entities (entity (person (sem_cat "SAM SKINNER"))))
    (sndrline (from (sem_cat "FROM:"))
       (entities (entity (person (sem_cat "EDE HOLIDAY"))))
    (subline (subj (sem_cat "SUBJECT:"))
       (topic (sem_cat "California Earthquake"))))
(body (paras (para
    (sem_cat "Attached is a situation report from FEMA on the northern California earthquake. No deaths have been reported and 45 people are known to have suffered injuries. In addition, there has been extensive property damage. While FEMA is awaiting a request from the State before initiating any recovery activities, a joint State/Federal preliminary damage assessment is likely to begin today."))
    paras (para
       (sem_cat "Director Stickney has requested that we forward the situation report to you.")))
    (optional (attachment (sem_cat "Attachments"))))
}
{qil=[document(e1),
document_form(e1, memo),
author(e1, 'EDE HOLIDAY'),
addressee(e1, 'SAM SKINNER'),
topic(e1, 'California Earthquake'),
date(e1, 'April 27, 1992')]}
```
A memorandum dated April 27, 1992 from EDE Holiday to Sam Skinner regarding California Earthquake.
Summary of Research Results & Current Research

• Results
  • A method for automatic document type recognition and metadata extraction.
  • Demonstrated it on 10 document types

• Current Research
  • Inducing grammars for documentary form from samples
  • Automatic Description of items, file units and record series
  • Automatic Recognition of the topics of records
Additional Information

Website: perpos.gtri.gatech.edu


Backup
Example of an Item Title and Description

Archival Description

56 results retrieved for President bush with filters applied

Refine Search  □ Highlight Search Terms  □ Return to Results

Description 50 of 56

Letter from George H. W. Bush to His Children on New Year’s Eve 1990, 12/31/1990
ARC Identifier 595134
Item from Collection GB-GBPP: George H. W. Bush Papers, ca. 1942 - ca. 2004

This letter was typewritten by President George H. W. Bush and addressed to his children: George, Jeb, Neil, Marvin, and Doro. He expresses his happiness at their Christmas celebration held at Camp David, then writes concerning his conflicted feelings as he prepares for the possibility of war with Iraq.
Archival Description Hierarchy: Series, File Unit, Item
Method for Extracting Person, Organization and Location Name Metadata

- Document Reader
- English Tokenizer
- Wordlist Lookup + Wordlists
- Sentence Splitter
- Hepple POS Tagger + Lexicon
- Named Entity Transducer + rules
- Extract Metadata
## Annotation and Metadata Extraction: Performance

<table>
<thead>
<tr>
<th>Annotation Type</th>
<th>Correct</th>
<th>Partially Correct</th>
<th>Missing</th>
<th>Spurious</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>515</td>
<td>11</td>
<td>42</td>
<td>57</td>
<td>0.8928</td>
<td>0.9164</td>
<td>0.9044</td>
</tr>
<tr>
<td>Location</td>
<td>270</td>
<td>15</td>
<td>54</td>
<td>24</td>
<td>0.8981</td>
<td>0.8186</td>
<td>0.8565</td>
</tr>
<tr>
<td>Organization</td>
<td>509</td>
<td>31</td>
<td>31</td>
<td>50</td>
<td>0.889</td>
<td>0.9186</td>
<td>0.9035</td>
</tr>
<tr>
<td>Date</td>
<td>456</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.9967</td>
<td>0.9967</td>
<td>0.9967</td>
</tr>
<tr>
<td>Money</td>
<td>28</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>0.7703</td>
<td>0.9828</td>
<td>0.8636</td>
</tr>
<tr>
<td>Percent</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Overall average precision: 0.9178  Overall average recall: 0.9282  Overall average F-measure: 0.9108
Annotated Document and Metadata Extracted For Indexing

<Date>April 27, 1992</Date>
<Date>7:00 a.m. EDT</Date>

SITUATION REPORT #3
PETROLIA EARTHQUAKE

DATE AND TIME
OF OCCURRENCE: <Date>April 25, 1992</Date>, <Time>11:06 a.m. EDT</Time>

LOCATION: <Location>Northern California</Location>, 30 miles southeast of <Location>Eureka</Location>

1. SITUATION:

   <paragraph>At <Time>2:06 p.m. EDT</Time> on <Date>April 25</Date>, a 6.9 Richter magnitude earthquake occurred in <Location>Northern California</Location>, 30 miles southeast of <Location>Eureka</Location> near the town of <Location>Petrolia</Location>.</paragraph>

   <paragraph>At <Time>3:42 a.m. EDT</Time> on <Date>April 26</Date> a strong aftershock (6.2 magnitude) occurred in the vicinity of <Location>Ferndale</Location> and <Location>Petrolia</Location> causing additional damage in <Location>Ferndale</Location>, <Location>Fortuna</Location> and <Location>Scotia</Location> -- but no report of additional injuries. The shock was felt as far away as <Location>San Francisco</Location>.</paragraph>

   <paragraph>At <Time>7:18 a.m. EDT</Time> on <Date>April 26</Date> a second aftershock registering 6.5 on the Richter scale occurred approximately 23 miles west of <Location>Cape Mendocino</Location>.</paragraph>

PERSON = (Nick Nikas, Roy Kite)

LOCATION = (Northern California, Eureka, Petrolia, Ferndale, Fortuna, Scotia, San Francisco, Cape Mendocino, Highway 254, Avenue of the Giants, San Andreas, Mendocino Fracture Zone, Cascadia Subduction Zone, Rio Dell, Route 101, Humboldt County, Fresno, Garberville)

ORGANIZATION = (Pacific Gas and Electric, PG&E, California Office of Emergency Services, FEMA, CAL OES, National Guard, Pacific Lumber Company, CALTRANS, Conservation Corps, Department of Water Resources, Air National Guard, Emergency Support Team, US Army Corps of Engineers, US Coast Guard, Salvation Army, Red Cross, Humboldt County Board of Supervisors)

A memorandum dated August 16, 1990 from Greg Petersmeyer to Nicholas Brady, Richard Jarman, and Michael Boskin regarding Charitable Deductions.

A memorandum dated September 18, 1990 from Ede Holiday to John Sununu regarding DOE's concerns on White House Process.

This file unit contains Cabinet Documents including memoranda relating to the World Bank Green Fund, Charitable Deductions and DOE's concerns on White House Process.
Series Description

This file unit contains Cabinet Documents including memoranda relating to the World Bank Green Fund, Charitable Deductions and DOE's concerns on White House Process.

This file unit contains materials relating to the 1992 Petrolia, California Earthquake. It includes memoranda, situation reports and correspondence.

This series consists of Cabinet Documents including memoranda relating to the World Bank Green Fund, Charitable Deductions and DOE's concerns on White House Process. This series also consists of memoranda, situation reports and correspondence relating to the 1992 Petrolia, California Earthquake.