Grammatical Induction and Recognition of the Documentary Form of Records

William Underwood, Sheila Isbell, Matthew Underwood
Georgia Tech Research Institute
250 14th St. N.W. Atlanta, Georgia 30332-0832
william.underwood@gtri.gatech.edu
sheila.isbell@gtri.gatech.edu
matthew.underwood@gtri.gatech.edu

Abstract

This paper presents digital curators with a more precise understanding of the concept of documentary form, and how documentary form can be automatically learned from a sample of records of a particular document type. The ability to automatically recognize documentary form enables item description. Item description enables file unit description and this enables automatic series description. This technology can reduce the effort required of an appraisal archivist to assess the value of record series containing a large number of e-records of different documentary forms. It can also provide archivists with earlier intellectual control of accessioned e-record series by providing preliminary scope and content notes for these series. Item descriptions provide additional ways for indexing and searching collections of records.

Introduction

Among the challenges archivists face in appraising e-records and gaining intellectual control of accessioned e-records is the enormous volume of records and the time it requires to read and understand the content of these records. According to one source, "the Clinton White House generated 38 million e-mail messages (and the current Bush White House is expected to generate triple that number)." [3] Archivists must review presidential records page-by page before they can be disclosed to the public or it is determined that here are restrictions on disclosure. Data collected on declassification review, indicates that a reviewer can review on average one page per minute, or 60 pages per hour. Given 1920 work hours per year, an archivist doing nothing other than review, could be expected on average to review 115,000 pages per year. NARA provides eight archivists to each Presidential Library, one of which is a Supervisory Archivist. Assuming seven archivists reviewing records, and an email with attachments averaging one page in length, they could review about 800,000 email massages per year. It will take 125 years for Presidential Library archivists to review and describe the Bush Administration's email for the first time.

In the next section, a method is described for recognizing the documentary form of records created by office applications such as word processors, spreadsheets and database management systems. Then it is shown how the ability to automatically recognize document type enables the automatic description of items, file units and record series. Finally, how these technologies can aid archivists in appraising e-records and gaining intellectual control of accessioned e-records is discussed.
.

Identifying the Documentary Form of E-Records

The science of Diplomatics defines the concept of documentary form as follows.

> "Diplomatics defines [documentary] form as the complex of the rules of representation used to convey a message, that is, as 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." *Intellectual form* 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." *Physical form* is "The overall appearance, configuration, or shape, derived from material characteristics and independent of intellectual content [1]."

Fig. 1 outlines a method for identifying the documentary form of a textual e-record of unknown type.

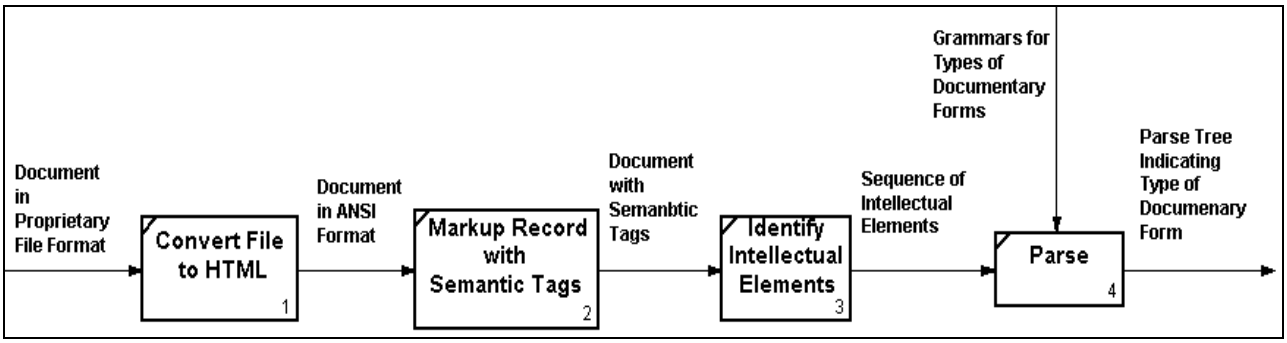

Figure 1. Identifying the Documentary Form of a Textual Document

Textual documents (word processing, spreadsheets or databases) in proprietary formats are converted into a common text format such as enriched text, HTML or PDF. Sentences are recognized and the text is annotated indicating paragraphs, tables and blocks of text such as postal addresses and lists. Semantic categories such as dates, person's names, organization names and locations are annotated. The sequence of intellectual elements of documentary form is identified. A parser uses grammars for a variety of documentary forms to identify the form of the document [4].

An Example

Fig. 2 shows a document that has been converted from a legacy proprietary file format to ANSI text with layout format.

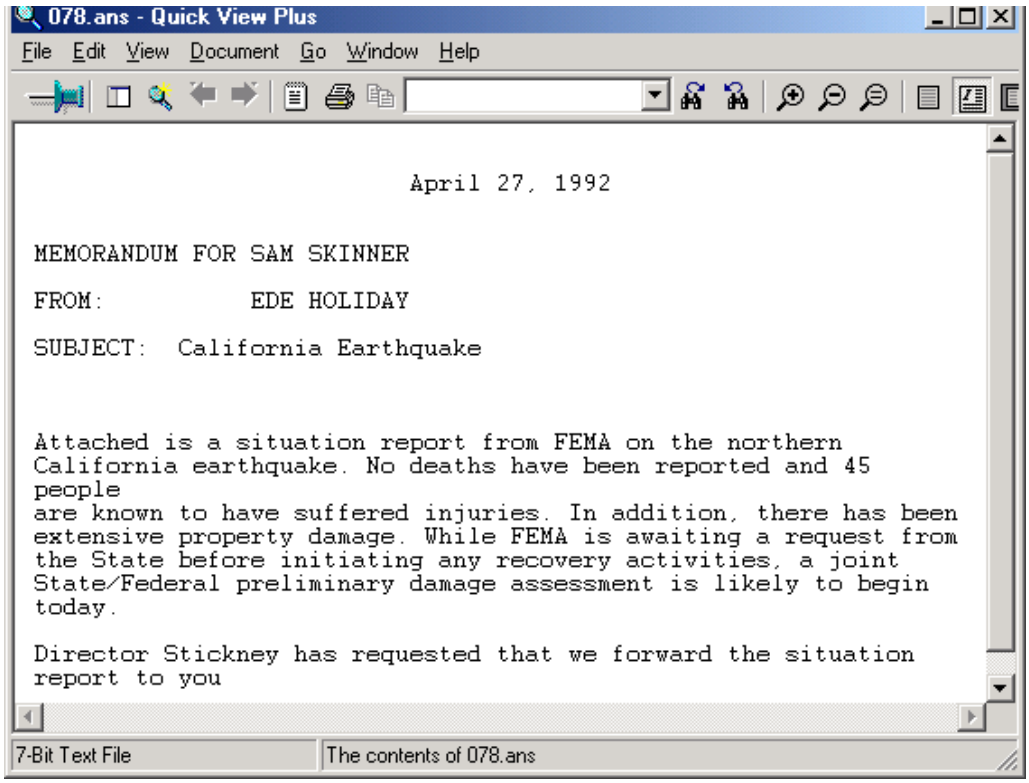

Figure 2. Document Converted to ANSI Text

Fig. 3 shows the ANSI text document annotated with semantic tags.

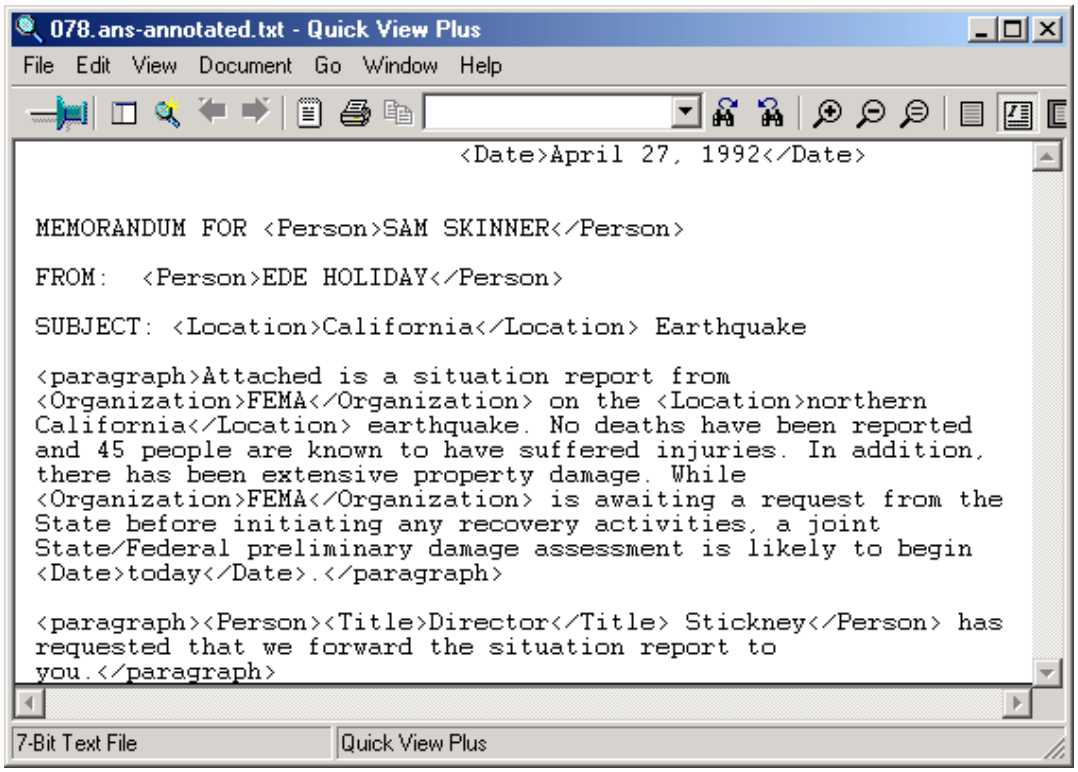

Figure 3. Document Annotated with Semantic Tags.

The Intellectual element identifier applies rules to remove the content leaving the intellectual form. It also associates those intellectual elements with their content. The intellectual elements for the document in Fig 3 are:

date MEMORANDUM FOR person FROM person SUBJECT np para para

The parser applies grammars for a number of documents types, e.g., memoranda, correspondence, resumes, press releases, schedules, and situation reports. Fig. 4 shows a simplified grammar for memoranda.

MEMO → HEAD BODY
HEAD → DATE FORPHRASE FROMPHRASE SUBJPHRASE
FORPHRASE → memorandum for PERSON
FROMPHRASE → from PERSON
SUBJPHRASE → SUBJ NP
BODY → PARAS
PARAS → PARAS PARA
PARAS → PARA
DATE → date
memorandum → MEMORANDUM
for → FOR
PERSON → person
from → FROM
SUBJ → SUBJECT
NP → np
PARA → para

Figure 4. A Simplified Context-Free Grammar for White House Memoranda.

The parser recognizes the sequence of intellectual elements as having the documentary form of a memo. The parse tree for the intellectual elements is shown in Fig. 5.

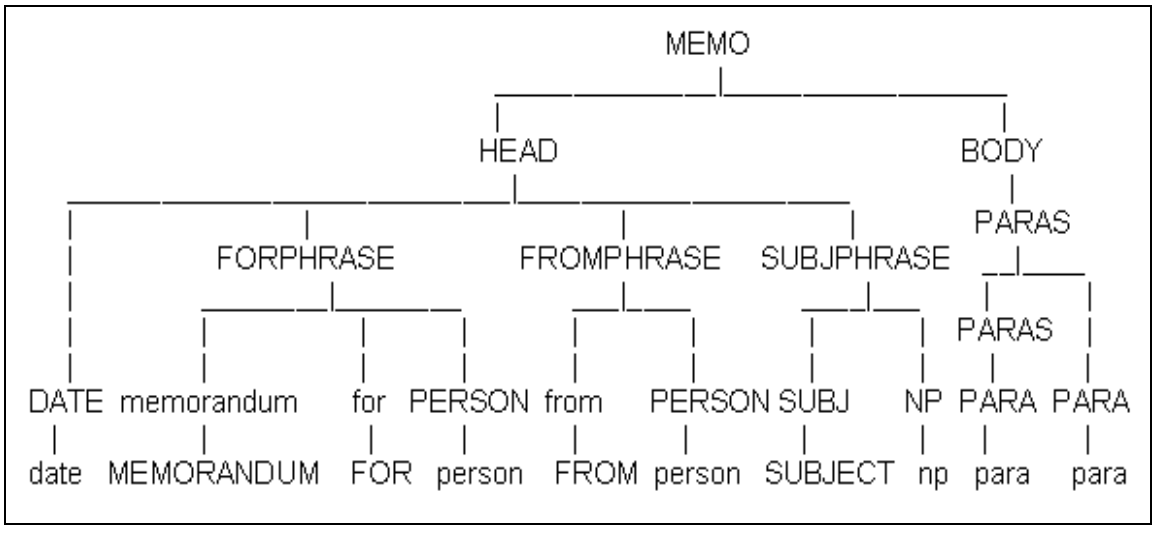

Figure 5. Parse Tree Indicating that the Document has the Form of a Memo.

Automatic Description of Items, File Units and Record Series

Item Descriptions

From the association of the content with the intellectual elements of documentary form, an item description can be automatically generated. For example, the following is the content associated with some of the intellectual elements of the memorandum shown in Fig. 2.

> Date = April 27, 1992
> For = SAM SKINNER
> From = EDE HOLIDAY
> Subject = California Earthquake

From the values of these elements, the following item description can be automatically generated.

> A memorandum dated April 27, 1992 from Ede Holiday to Sam Skinner regarding California Earthquake.

File Unit Descriptions

Suppose that one of the file units (folders, directories) in *Edith E. Holidays Files* is *Cabinet Documents*, and in this file unit, three memos were identified and their intellectual elements and content used to create the following item descriptions.

> A memorandum dated June 7, 1990 from John Niehuss to Stephen Janzansky regarding World Bank Green Fund.
>
> 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

From these item descriptions the following file unit description can be automatically generated.

> 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 Descriptions

Suppose that there was another file unit in *Edith E. Holiday's Files* titled *Petrolia, Calif. (Cape Mendocino) Earthquake* and that its automatically generated description is

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

Then the following series description could be automatically generated for *Edith E. Holiday's Files*.

> 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.

These descriptions are not as informative as those that are written by archivists, but they can provide archivists with information about the content of record series long before archivists have found the time to read the individual records.

Inducing a Grammar for the Documentary Form of a Document type

One doesn't have to manually construct grammars for document types. Information extraction and grammatical induction technology are being applied to the problem of learning the documentary form of a variety of Presidential electronic records. Given a sample of records of a particular type, such as correspondence, memoranda or agenda, file format conversion technology is used to convert files in proprietary formats to a standard text format such as enriched text or HTML. Then information extraction (semantic tagging) technology is used to identify and annotate semantic categories of terms, such as person's names, organization names, location names, communication acts, job titles, dates, and postal addresses [2]. The intellectual elements of documentary form are then identified. From the intellectual elements of the sample records, a stochastic context-free grammar is automatically induced that defines the documentary (intellectual) form of that particular document type. Grammars learned for a variety of record types can then be used with a parser to recognize documentary forms of records of unknown document type [4].

The method of grammatical induction can improve (learn) the grammar for a document type when it is provided additional samples. Furthermore, the method of manually constructing a grammar for a document type can be combined with the method of automatic grammatical induction to prefer the phrase structure prescribed by the creator of the grammar. This enables automatic improvement of the grammar in the face of variations in the documentary form due to authors of documents not always following stylistic guidelines to the letter.

Use of these Tools in Appraisal, Accession, Review, Description, Search and Retrieval

Selection and Appraisal

Due to large volumes of e-records in record series, an appraisal archivist may only review a sample of the records. These tools can assure more complete assessment when appraising large record series, by identifying all the document types and summarizing the entire series.

Accession

When e-records that have been scheduled for long-term preservation are transferred to archive, it should be verified that the records received are the ones that are expected. The capability to recognize the documentary form and action or matter of the records can be part of this verification.

Review

Knowing a record's type and participants aids in understanding the action communicated by a record, for example, a correspondence record type between family members may indicate

personal communications. Thus, knowing the action communicated by a record can aid in discriminating personal records from Presidential records. A decision memo record type regarding an appointment communicates an appointment action. Some records related to Presidential appointments to Federal Office are restricted from disclosure under the Presidential Records Act. Memoranda that provide advice or recommendations may be subject to PRA restriction a(5) that restricts for a period of 12 years after the end of an administration the disclosure of confidential advice between the President and his advisors.

Description

Document types are elements of archival descriptions -- item, file unit and record series descriptions. As described in an earlier section of this paper, from knowing the document type of a record, one can automatically extract metadata such as chronological date, author and addressee, and sometimes action or matter. Having used this information to describe the items in a directory, one can automatically describe that file unit. Having described the file units in a record series, one can automatically generate a description of the record series.

Search and Retrieval

The ability to recognize documentary form can also contribute to searching for and retrieving records. For instance, having determined the documentary form of records in a repository, and having the elements of the form annotated, one can index the records on these elements, e.g. document type, author, date, addressee, and action or subject. Then one can constrain the search for records to be based on these elements.

Research Issues

Context-free grammars have been created for a score of document types that occur in the White House Office Staff Member and Office Files of the George H. W. Bush Presidential E-records. These include Briefing Memos, Decision Memos, Information Memos, Referral Memos, Signature Memos, Telephone Call Recommendations, FEMA Situation Reports, Executive Orders, National Security Directives, White House Correspondence, Schedules, Mailing Lists, Database Tables, and Spreadsheets. There are at least 30 additional document types that occur in this collection whose form needs to be defined.

The grammatical induction method requires a sample of 100 or more e-records in order to generalize a grammar that is comparable to those generated by a person. Samples of the document types contained in the Bush Presidential e-records are currently being gathered for use in inducing the documentary forms of these document types.

The recognition of the subject of memoranda is relatively easy as the subject is explicitly specified. The recognition of the subject or topic of reports is usually easy in that the title usually indicates the subject. However, many document types such as correspondence may not have a subject line or title and it is necessary to analyze the content of the document to determine the subject or action performed by the record (e.g., resignation, appointment, advice, or recommendation).

The methods for inducing a grammar and recognizing the documentary form of textual e-records are currently limited to the induction and recognition of intellectual form. The methods for identifying intellectual elements need to be extended to identify physical elements of form as well.

Acknowledgements

The ERA Program of the National Archives and Records Administration and the Army Research Laboratory sponsored this research under Cooperative Agreements DAAD19-03-2-0018 and W911NF-06-2-0050

References

1. Duranti. *Diplomatics: New Uses for an Old Science.* Society of American Archivists and Association of Canadian Archivists in association with Scarecrow Press, 1998.
2. S. Isbell and M. Underwood and W. Underwood. The PERPOS Information Extractor Applied to Presidential E-Records. PERPOS TR ITTL/CSITD 05-10, November 2006
3. D. Talbot. The Fading Memory of the State, *Technology Review*, July 2005
4. W. E. Underwood and B. Harris. Inferring and Recognizing the Documentary Form of Record Types. PERPOS TR ITTL/CSITD 05-8, August 2006.