Inferring and Recognizing the Documentary Form of Record Types

William E. Underwood
Brian Harris

PERPOS Technical Report ITTL/CSITD 05-08

August 2006

Computer Science and Information Technology Division
Information Technology and Telecommunications Laboratory
Georgia Tech Research Institute
Georgia Institute of Technology

The Army Research Laboratory (ARL) and the National Archives and Records Administration (NARA) sponsor this research under Army Research Office Cooperative Agreement DAAD19-03-2-0018. The findings in this paper should not be construed as an official ARL or NARA position unless so indicated by other authorized documentation.

Abstract

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, information extraction technology is used to identify and markup semantic categories, such as person's names, organization names, location names, job titles, dates, and postal addresses. The content is then removed from the annotated records leaving the intellectual elements of documentary form. From the intellectual elements of the sample records, a stochastic context-free grammar is automatically induced that defines the documentary form of that particular record type. Grammars learned for a variety of record types can then be used with a parser to recognize documentary forms of records of unknown record type.

The significance of this research is that record types have a role in archival description and review. Archival descriptions include the names of the types of records that occur in a record series. In addition, knowing a record's type aids in understanding the action communicated by a document. Knowing record type can aid in discriminating personal records from Presidential records. It can also aid in determining access restrictions. The ability to recognize record type can also contribute to searching for and retrieving records based on the intellectual elements of their documentary form.

Keywords: record type, document type, documentary form, grammatical induction

Table of Contents

1. INTRODUCTION 1

1.1 BACKGROUND 1
1.2 PURPOSE 1
1.3 SCOPE 1

2. DOCUMENTARY FORM AND DOCUMENT TYPE 2

2.1 DOCUMENTARY FORM 2
2.2 MARKUP LANGUAGES AND DOCUMENT TYPES 3

3. METHOD FOR INFERRING THE DOCUMENTARY FORM OF RECORDS .. 5

3.1 THE RESEARCH PROBLEM 6
3.2 AN APPROACH 7
3.2.1 Information Extraction 8
3.2.2 Content Removal 9
3.2.3 Induction of Stochastic Context-Free Grammars 10
3.3 INFERRING GRAMMARS FOR THE DOCUMENTARY FORM OF RECORD TYPES 12
3.3.1 A Grammar for the Documentary Form of White House Correspondence 13
3.3.2 A Grammar for the Documentary Form of White House Memoranda 17

4. METHOD FOR RECOGNIZING DOCUMENTARY FORM 22

5. RELATED RESEARCH 24

6. RESULTS AND RESEARCH ISSUES 26

REFERENCES 28

APPENDIX A: SAMPLE CORRESPONDENCE AND MEMORANDA RECORD TYPES 31

List of Figures

Figure 1. An XML Document 4
Figure 2. External DTD for the Memorandum 5
Figure 3. Record Types in the Bush Administration's Personal Computer Files 6
Figure 4. Method for Inducing the Documentary Form of Record Types. 7
Figure 5. XML Annotations for Named Entities in a Document 8
Figure 6. Sample White House Correspondence. 13
Figure 7. XML Markup of Named Entities in Sample Correspondence. 14
Figure 8. Intellectual Forms of Sample White House Correspondence 15
Figure 9. Induced SCFG for the Correspondence Record Type. 16
Figure 10. A Better Grammar for White House Correspondence. 17
Figure 11. Sample White House Memorandum. 18
Figure 12. XML Markup of Named Entities in the Sample Memo 19
Figure 13. Intellectual Forms of Sample White House Memoranda 20
Figure 14. A Context-Free Grammar Induced from Sample White House Memoranda .. 21
Figure 15. A Better Grammar for White House Memoranda 22
Figure 16. Method for Recognizing Documentary Form. 23
Figure 17. Parse Showing the Documentary Form of a Record. 23
Figure 18. Graphical Parse Tree Showing Documentary Form of a Record 24

1. Introduction

1.1 Background

The capability to recognize the documentary form of a record is an important element of archival description. Archival descriptions include the names of record types such as correspondence, memoranda, and agenda [Underwood 2005].

The recognition of documentary form is also important in identifying the authors and addressees of records, and the speech or communication act conveyed by a record. Speech acts include such acts as resigning, appointing, nominating, advising, recommending, requesting, briefing, reporting and many other human actions that are carried out in Presidential records. This information is useful in determining whether Presidential Record Act restrictions might apply to Presidential records [Underwood and Harris 2005].

The ability to recognize documentary form can also contribute to retrieval of relevant records from a collection. Having recognized the form of a record and its intellectual elements, these can be used to index the records and provide the user with the capability to retrieve just those with specific forms or elements [Iwanska and Underwood 2006].

1.2 Purpose

The purpose of this paper is to describe a method for automatically inferring or inducing the documentary structure of types of Presidential e-records and using the grammars to recognize the documentary structures of records of unknown record type.

1.3 Scope

In section 2, the concept of documentary form as studied by the science of Diplomatics and the concept of document type as specified in the Extensible Markup Language (XML) are reviewed. In section 3, a method for inducing grammars for the documentary forms of record types is described. The method has been implemented and this section gives two examples of its application to Presidential e-records. Section 4 describes the method of using the grammars with a parser for identifying or classifying record types. In section 5, related work by computer scientists in recognizing documentary form and in grammatical inference is reviewed. In section 6, results are summarized and research issues identified.

2. Documentary Form and Document Type

2.1 Documentary Form

Diplomatics is the study of the creation, forms, and transmission of records, and their relationship to the facts represented in them and to their creator, in order to identify, evaluate, and communicate their nature and authenticity. The science of Diplomatics originated in the need to authenticate medieval documents, but has been extended to address the nature and authenticity of modern documents including electronic records.

Documentary form is "the rules of representation according to which the content of a document, its administrative and documentary context, and its authority are communicated [InterPARES 2001b]. The form of a document derives from the business activity and the formal (or informal) procedure used to create it.

Documentary form consists of physical form and intellectual form. The *physical form* of a document is the overall appearance, configuration, or shape, derived from its material characteristics and independent of its intellectual content. 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 term *physical form* refers to the external make-up of the document, while the term *intellectual form* refers to its internal articulation. Therefore, the elements of the former are defined as external or *extrinsic*, while the elements of the latter are defined as internal or *intrinsic* [Duranti 1998, p 134]. Extrinsic elements include medium, script, language, special signs, seals, and annotations. "Extrinsic elements refer to specific, perceivable features of the record that are instrumental in communicating and achieving the purpose for which it was created [InterPARES 2001a, p. 5]. For electronic records, these include:

- overall presentation features (e.g., textual, graphic, image, sound, or some combination of these);
- specific presentation features (e.g., special layouts, hyperlinks, colors, sample rate or sound files);
- electronic signatures and electronic seals (e.g., digital signatures);
- digital time stamps;
- other special signs (e.g., digital watermarks, an organization's crest or personal logo).

Intrinsic elements are the discursive parts of the record that communicate the action in which the records participates and the immediate context. They fall into three groups:

1) elements that convey aspects of the record's juridical and administrative context (e.g., the name of the author, addressee, the date);

2) elements that communicate the action itself (e.g., the indication and description of the action or matter);
3) elements that convey aspects of the record's documentary context and its means of validation (e.g., the name of the writer, the attestation, the corroboration). [InterPARES 2001a, p. 5]

2.2 Markup Languages and Document Types

Markup originated in the publishing industry. In traditional publishing, a copy editor annotates a manuscript with layout instructions for the typesetter. These handwritten instructions are called markup.

Word processing applications, such as WordPerfect, require that the user specify the appearance of the text. The user selects commands in menus to add formatting instructions to the text. For instance, one might select the title, and then select 16pt, Arial, center-aligned. These formatting instructions control how the document will be displayed or printed. This type of markup is often referred to as specific markup. This markup is conceptually similar to the handwritten instructions for the typesetter.

Many word processing applications, such as WordPerfect and MSWord, simplify this formatting process with sets of macros called styles. A style is a set of formatting characteristics that you can apply to text in your document to quickly change its appearance. For example, to format the title of a report, instead of taking three separate steps to format a title as 16 pt, Arial, and center-aligned, you can achieve the same result in one step by applying the Title style. The title style or macro generates these three instructions. This kind of markup is referred to a generic markup.

The Standard Generalized Markup Language (SGML) [ISO 1986] is similar to generic markup. SGML is a language to describe documents. The markup describes the document's structure, not the document's appearance. Document structure is specified in a Document Type Definition (DTD), sometimes referred to as a SGML application. A DTD specifies a set of elements, their relationships, and the tag set to markup the document.

Hypertext Markup Language (HTML) is a SGML application for publishing documents on the World Wide Web. HTML has been defined as an SGML DTD [W3C 1999a].

SGML and HTML are very complex markup languages. The Extensible Markup Language (XML) is a simpler subset of SGML that retains it essential features [W3C 2006a]. Figure 1 shows a record from the Bush Presidential Records that has been marked up as an XML document.[1]

[1] Bush Presidential Library, Bush Presidential Records, WHORM Subject File, Disasters-Natural, ID#324869.

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE memo SYSTEM "http://www.site.com/dtds/memo.dtd"
<memo>
            <header>

<date>April 27, 1992</date>

<for type="MEMORANDUM FOR">
            <person>SAM SKINNER</person>
</for>
<from type="FROM:">
            <person>EDE HOLIDAY</person>
</from>
<subject type="SUBJECT:">
            <np>California Earthquake</np>
</subject>
</header>
<body>
<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>
</body>
</memo
```

Figure 1. An XML Document

The second line of the XML document is a Document Type Declaration. It links the document file to a Document Type Definition (DTD). The structure of the document is described by pairs of XML tags, for example, <date> April 27, 1992</date> that bracket content.

XML DTDs specify the structure of XML documents. Figure 2 shows the DTD for the memorandum in Figure 1. The DTD specifies that a memo consists of a header element followed by a body element. The header consists of a sequence of date, for, from and subject elements. The body consists of a sequence of one or more paragraphs.

```
<!DOCTYPE memo [
<!ELEMENT       memo      (header, body)                                  >
<!ELEMENT       header    (date, for, from, subject)                      >
<!ELEMENT       date      (#PCDATA)                                       >
<!ELEMENT       for       (person)                                        >
<!ATTLIST       for       type      NMTOKENS     "MEMORANDUM FOR"       >
<!ELEMENT       person    (#PCDATA)                                       >
<!ELEMENT       from      (person)                                        >
<!ATTLIST       from      type      NMTOKEN        "FROM:"               >
<!ELEMENT       subject   (np)                                            >
<!ATTLIST       subject   type      NMTOKEN        "SUBJECT:"            >
<!ELEMENT       np        (#PCDATA)                                       >
<!ELEMENT       body      (para+)                                         >
<!ELEMENT       para      (#PCDATA)                                       >
]>
```

Figure 2. External DTD for the Memorandum

XML is used to structure the information in documents apart from its appearance. To view or print XML documents, they must be formatted or styled. Style sheets specify how XML documents should be rendered when displayed on the screen or in an editor, or printed on paper. XML is supported by two style sheet languages--XML Stylesheet Language (XSL) [W3C 1999b] and Cascading Style Sheet (CSS) [W3C 2006b].

An XSLT stylesheet specifies the presentation of a class of XML documents by describing how an instance of the class is transformed into an XML document that uses a formatting vocabulary, such as XHTML [W3C 2000] or XSL-FO [W3C 2001]. The style sheet consists of a set of rules (templates) that match the XML elements of a document and associate them with a list of elements in order to format them.

The concept of document structure defined by a XML DTD is a formal model of the Diplomatics concept of the intellectual form of a document. The concept of an XSL stylesheet is a formal model of the concept of the physical form of a document.

3. Method for Inferring the Documentary Form of Records

Archivists use the term record type, rather than document type. There are two related senses of the term [Pearce-Moses 2004].

1. A distinctive class of records defined by their function or use.
2. A class of records defined by their style, subject, physical characteristics, or form.

Examples of record types in the first sense include baptismal records, deeds, and accounting ledgers. Examples of record types in the second sense include moving images, photographs, and oral histories.

A related term is *genre* [Pearce-Moses 2004]

> A distinctive type of literary or artistic materials, usually characterized by style or function rather than subject, physical characteristics, or form.
>
> Typical examples of genres include correspondence and contracts. The subject matter and form of a genre may be quite varied. For example, a contract may relate to work done for hire, the loan of materials, or purchase of materials. Those contracts may be take the form of personalized letters or boilerplate documents and may be preprinted forms or holographs.

Fig. 3 shows some of the record types that occur in the personal computer files of staff members of the Presidential Administration of George H. W. Bush.

Agenda	Newsletter
Attendee List	Nomination to Federal Office
Bar Chart	Notes
Biography	Presidential Statement
Briefing (Presentation)	Press Pool Report
Briefing Memo	Press Release
Correspondence	Proclamation
Decision Memo	Recipe
Diary	Referral Memo
Executive order	Resume
Information Memo	Schedule
Job Application	Signature Memo
Letter	Situation Report
List of Candidates	Staff Register
Mailing List	Summary
Memo	Telephone Call Recommendation
Minutes	Transcript of News Conference
National Security Directive (NSD)	Transcript of Speech

Figure 3. Record Types in the Bush Administration's Personal Computer Files

NARA's *Life-Cycle Data Requirements Guide* specifies that in creating the Scope and Content Note for a record series, an archivist should "Enter the information about the specific records types, such as reports, minutes, correspondence, speeches, questionnaires, drawings or photographs.

3.1 The Research Problem

Personal computer records are created in many proprietary file formats, most of which use specific markup to indicate the physical form or style of the records. Archivists are able to easily recognize record types and use this knowledge in describing record series.

If it were possible to automatically recognize a record's type, it would enable automatic description of record series. If one could automatically recognize the documentary structure of a record, one could annotate (tag) a copy of the record showing its structure. This could facilitate understanding the action carried out by the record and its participants, both important aspects of a record for determining whether it might have access restrictions. In addition, the annotated copy could be used for indexing the records according to their author, recipient, subject or matter, and date that can be important features in formulating queries used to search repositories of records for records relevant to a FOIA request.

This research seeks to answer these research questions:

> Given a sample of e-records of a particular record type, e.g., memoranda or correspondence, can a method be developed and demonstrated to automatically learn the documentary form of records of that type?
>
> Given a textual e-record, can a method be developed and demonstrated to automatically recognize its documentary form and thus identify its record type?

An approach to answering the first research question is described in this section of the report. An approach to answering the second question is discussed in section 4.

3.2 An Approach

Figure 4 outlines a method for using grammatical induction for inferring the documentary form of record types.

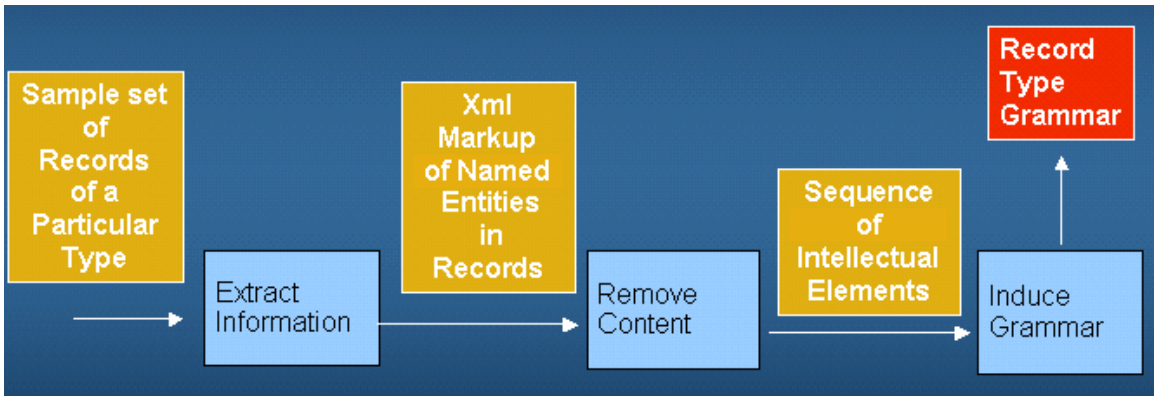


Figure 4. Method for Inducing the Documentary Form of Record Types.

A set of e-records of a particular record type is selected and converted to text or html format. An information extractor is applied to this sample to annotate the named entities in the records. An algorithm is applied that removes the content from the annotated records leaving a list of the intellectual elements in each record. Grammatical induction is applied to these lists to produce a stochastic context-free grammar for the intellectual form of the documentary form of the record type. The following sections provide the details of the method.

3.2.1 Information Extraction

Information extraction (IE) is a technology for identifying semantic categories in text and annotating them to produce structured information such as marked-up text, templates or database tables. The ANNIE Information Extractor has been refined to reliably recognize dates, person's names, organization names, location names, job titles, and postal addresses [Isbel and Underwood 2006].

Figure 5 shows the result of applying the Information Extractor to the document shown in Figure 1.

```
                                        <paragraph><Date>April 27,
1992</Date>
</paragraph>

<paragraph>MEMORANDUM FOR <Person>SAM SKINNER</Person>
</paragraph>
<paragraph>FROM:          <Person>EDE HOLIDAY</Person>
</paragraph>
<paragraph>SUBJECT:       <Location>California</Location> Earthquake
</paragraph>


<paragraph>Attached is a situation report from
<Organization>FEMA</Organization> on the northern
<Location>California</Location> earthquake. No deaths have been reported and 45
people are known to have suffered injuries. In addition, there
has been extensive property damage. While <Organization>FEMA</Organization> is
awaiting a
request from the State before initiating any recovery activities,
a joint State/Federal preliminary damage assessment is likely to
begin <Date>today</Date>.
</paragraph>
<paragraph><Person><Title><JobTitle>Director</JobTitle></Title>
Stickney</Person> has requested that we forward the situation
report to you.
</paragraph>
```

Figure 5. XML Annotations for Named Entities in a Document

Early in the information extraction process, segments of text separated by white space are tagged with pairs of <paragraph> tags. Dates, person's names, organization names, location names, titles, and job titles are similarly tagged. While XML tags are used to

annotate the named entities appearing in the document, the document is not an XML document. It has no Document Type Definitions and the Document elements are not hierarchically structured.

3.2.2 Content Removal

An algorithm has been constructed that removes the content from records that have been annotated by the Information Extractor. The remaining markup represents the intellectual elements of the documentary form of the records.

The following rules are used for content removal.

1. <paragraph>*sentence*$^{+}$</paragraph> → paragraph
2. <paragraph><date>*content*</date></paragraph> → date
3. <paragraph><person>*content*</person></paragraph> → person
4. <paragraph><organization>*content*</organization></paragraph> → organization
5. <paragraph><address>*content*</address></paragraph> → address
6. <paragraph><jobtitle>*content*</jobtitle></paragraph> → jobtitle
7. <paragraph>*string*<X>*content*</X></paragraph> → *string*X
8. <paragraph>*CapitalLetter. nonsentence*</paragraph> → heading
9. <paragraph>*DecimalNumeral. nonsentence*</paragraph> → heading
10. <paragraph>*SmallLetter. nonsentence*</paragraph> → heading

The information extractor includes a text segmenter that annotates all blocks of text with the XML elements <paragraph>*text*</paragraph>, even though the *text* may not be a sequence of sentences. Only a sequence of sentences will be considered to be the intellectual element *paragraph* as seen in the first rule. In that rule, the plus (+) superscript on *sentence* indicates one or more sentences.

Rules 2-6 remove the specific dates, person's names, organization names, addresses and jobtitles, and replace the XML markup with the intellectual elements date, person, organization, address, and jobtitle, respectively.

Rule 7 applies to markup such as

<paragraph>To: <person>Governor Sununu</person></paragraph>

to produce

To: person.

Terms and punctuation such as "Memorandum," "To:" and "From:" that appear in memoranda are also intellectual elements.

Rules 8-10 identify section headings appearing in memoranda and reports. Nonsentences are strings of characters that are not parseable as a sentence and do not end in a period, question mark, or exclamation point.

3.2.3 Induction of Stochastic Context-Free Grammars

Grammatical inference (or induction) is a technique for discovering structure in input data. It is an inductive inference task that takes a set of sample strings and returns a description of their common structure.

Stolcke and Omohundro describe a grammar induction algorithm that performs a beam search through space of stochastic context-free grammars (SCFGs), guided by a Bayesian evaluation function [Stolcke and Omohundro, 1994]. Their algorithm generates candidate grammars using a chunking operator that creates new nonterminals for repeated substrings and a merging operator that combines nonterminals. The evaluation function decomposes the prior probability of a candidate into a product of two components: structure, calculated by description length; and parameters (rule probabilities), calculated by a Dirichlet distribution. Posterior probabilities are estimated on structure alone, by using the Viterbi approximation to integrate over all possible parameter settings.

An expectation maximization procedure is used to achieve a locally optimal grammar. Given a body of data X, the model merging algorithm generates a model M that maximizes the P(X|M). Then it generates new models by applying generalization/merging operators that combine rules and their associated probabilities. A search method is used to find the model that best maximizes the posterior probability P(M|X) according to Bayes' Rule.

A *context-free grammar* (CFG) is formally defined as a 4-tuple M = <V, Σ, R, S> where

V is a set of *non-terminals symbols*.
Σ is a set of *terminal symbols* (or alphabet) that is disjoint from V.
R is a set of *production rules* of the form $X \rightarrow y$, Where $X \in V$ and *y* is a string of finite length that it made up of symbols in V and Σ.
$S \in V$ is called the *initial symbol*.

A stochastic context-free grammar (SCFG) is a context-free grammar that associates a probability with each rule in the grammar. The probabilities of rules with the same nonterminal on the left-hand-side sum to zero. The primary advantage of a SCFG over a CFG is that it can help disambiguate sentences recognized by a parser using the SCFG.

Stolke's algorithm for grammatical induction begins with a grammar constructed by

1. Creating a nonterminal symbol producing each terminal in the sample.
2. Creating rules with the symbol S on the left-hand side and the string of nonterminals replacing each terminal in a list on the right hand side.

For instance, if the language sample is:

ab
aabb
aaabbb

the initial grammar would be

$S \rightarrow AB$
$S \rightarrow AABB$
$S \rightarrow AAABBB$
$A \rightarrow a$
$B \rightarrow b$

3. Subsequent models are generated using two operators called chunking and merging.

Chunking is substituting a nonterminal for a sequence of two or more nonterminals that occur on the right-hand-side of two or more productions and adding a production with the new nonterminal on the left-hand side and the sequence of two or more nonterminals on the right-hand side. In the initial grammar above, the string AB appears on the right-hand side of three productions. The Chunking operator replaces AB with X and adds the production $X \rightarrow AB$

$S \rightarrow X$
$S \rightarrow AXB$
$S \rightarrow AAXBB$
$X \rightarrow AB$
$A \rightarrow a$
$B \rightarrow b$

The chunking operator alone would generate simpler grammars, but never more than exactly the corpus. The merging operator supports generalization of the grammar. The merging operator replaces two nonterminals X_1 and X_2 with a single nonterminal Y, which may also be either X_1 or X_2. It is written as merge(X_1, X_2) = Y. It is defined as follows:

- Right-hand side occurrences of X_1 and X_2 are replaced by Y
- Nonterminals X_1 and X_2 on the left-hand side of productions are replaced by Y

Consider the grammar below.

S → X
S → Y
S → AYB
X → AB
Y → AXB

The merging operator merge(Y, S) =S replaces Y in all productions with S.

S → X
S → S
S → ASB
X → AB
S → AXB

Whenever merging produces a rule of the form Z→ Z, it is eliminated from the grammar. So, S → S is eliminated.

4. Derivative models log likelihoods are determined. The one with the highest likelihood is selected. If the derived grammar with the highest likelihood is has a higher score than I_0 then set the grammar equal to the new I_0 and execute step 3 again. If there are no derived grammars that score higher then the initial model then, depending on the search mechanism, the initial model can be returned or a search that looks a few states ahead can be used to determine if in x number of steps a model can be produced that scores higher then the initial model.

After applying the grammatical induction method to this example, the following grammar is selected.

S → A B
S → A S B
A → a
B → b

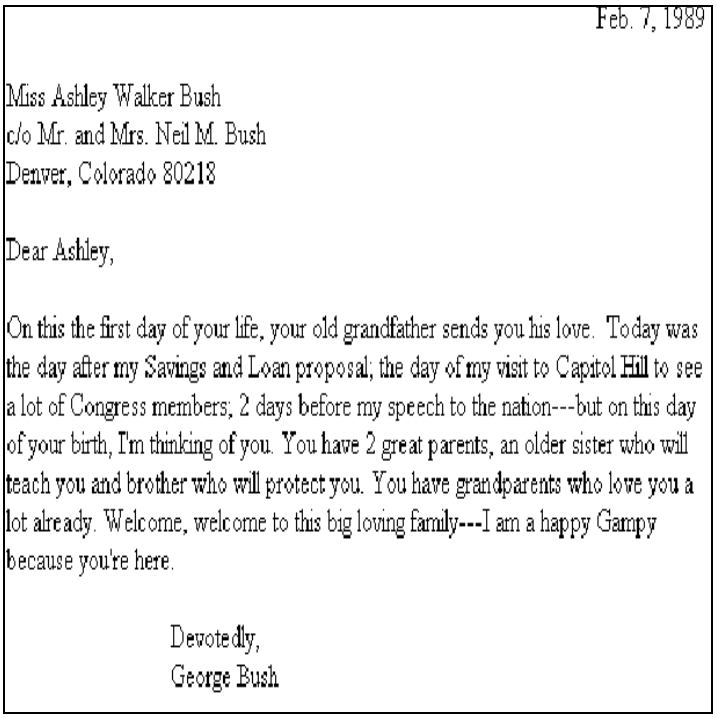
It describes the structure of the strings $a^n b^n$, that is one or more a's followed by the same number of b's.

3.3 Inferring Grammars for the Documentary Form of Record Types

In this section, examples are given of applying the record type induction method to White House correspondence and memoranda.

3.3.1 A Grammar for the Documentary Form of White House Correspondence

The grammatical induction method was applied to a sample of White House correspondence (See Appendix A). The record shown in Figure 6 is in the sample.[2]



Feb. 7, 1989

Miss Ashley Walker Bush
c/o Mr. and Mrs. Neil M. Bush
Denver, Colorado 80218

Dear Ashley,

On this the first day of your life, your old grandfather sends you his love. Today was the day after my Savings and Loan proposal; the day of my visit to Capitol Hill to see a lot of Congress members; 2 days before my speech to the nation---but on this day of your birth, I'm thinking of you. You have 2 great parents, an older sister who will teach you and brother who will protect you. You have grandparents who love you a lot already. Welcome, welcome to this big loving family---I am a happy Gampy because you're here.

Devotedly,
George Bush

Figure 6. Sample White House Correspondence.

The information extractor was applied to this sample to produce annotations that describe the named entities in the records. Figure 7 shows the XML annotations for the named entities in the record above.

[2] *All the Best, George Bush: My Life in Letters and other Writings*, Scribner, 1999, p. 413.

```
<paragraph><Date>Feb. 7, 1989</Date></paragraph>

<paragraph><Address><Person><Title>Miss</Title> Ashley Walker
Bush</Person>
c/o <Title>Mr.</Title> and <Person><Title>Mrs.</Title> Neil M. Bush</Person>
<City><Location>Denver</Location></City>,
<State><Location>Colorado</Location></State>
<Location>80218</Location></Address></paragraph>

<paragraph>Dear <Person>Ashley</Person>,</paragraph>

<paragraph>On this the first day of your life, your old grandfather sends you his love.
<Date>Today</Date> was the day after my Savings and Loan proposal; the day of
my visit to <Location>Capitol Hill</Location> to see a lot of
<Organization>Congress</Organization> members; <Date>2 days</Date> before
my speech to the nation---but on this day of your birth, I&apos;m thinking of you.
You have 2 great parents, an older sister who will teach you and brother who will
protect you. You have grandparents who love you a lot already. Welcome, welcome
to this big loving family---I am a happy Gampy because you're here.</paragraph>

                    <paragraph>Devotedly,
                <Person>George Bush</Person>
</paragraph>
```

Figure 7. XML Markup of Named Entities in Sample Correspondence.

The content removal algorithm is applied to the annotated records leaving a list of the intellectual elements of documentary form for each record in the sample. Figure 8 shows the intellectual elements of the documentary form of twenty-one pieces of correspondence. The third list in the figure corresponds to the intellectual elements in the record shown in Figs 6 and 7.

(DATE Dear PERSON PARAGRAPH PARAGRAPH Warmly ADDRESS);;C:\testDir\001.txt
(DATE ADDRESS Dear PERSON PARAGRAPH Devotedly PERSON);;C:\testDir\005.txt
(DATE ADDRESS Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON);;C:\testDir\007.txt
(DATE ADDRESS Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH With best regards Sincerely PERSON);;C:\testDir\008.txt
(DATE ADDRESS Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH All Best PERSON);;C:\testDir\009.txt
(DATE ADDRESS Dear PERSON PARAGRAPH PARAGRAPH Sincerely PERSON);;C:\testDir\016.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH With best regards Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\029.txt
(Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH With best regards Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\031.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\037.txt
(DATE ADDRESS Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE);;C:\testDir\059.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\068.txt
(DATE ADDRESS Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE);;C:\testDir\070.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH With best regards Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\072.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH With best regards Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\076.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ORGANIZATION ADDRESS);;C:\testDir\080.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\082.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\084.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\086.txt
(DATE Dear PERSON PARAGRAPH Warmly PERSON ADDRESS);;C:\testDir\106.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH PARAGRAPH Warmly PERSON ADDRESS);;C:\testDir\107.txt
(DATE Dear PERSON PARAGRAPH PARAGRAPH Sincerely PERSON JOBTITLE ADDRESS);;C:\testDir\112.txt

Figure 8. Intellectual Forms of Sample White House Correspondence.

To induce a stochastic context-free grammar for the intellectual form of White House correspondence, Stolke's algorithm for grammatical induction was applied to this sample. The result is shown in Figure 9.

```
NT2219    (NT2109 NT2109)                          1.000    ;43.0000
NT2239    (NT2219 NT2219 NT2219)                   1.000    ;12.0000
NT2692    (NT2692 NT2219)                          0.325    ;7.00000
          (NTDATE-7896-2081 NT2083 NT2084)         0.675    ;14.0000
NT2709    (NT2109 NT2086 NT2084)                   0.349    ;13.0000
          (NT2109 NT2709)                          0.120    ;5.00000
          (NT2121 NT2084)                          9.143e-2;4.00000
          (NT2239 NT2709)                          0.206    ;8.00000
          (NT2709 NT2087)                          0.234    ;9.00000
S       (NT2692 NT2121 NT2082)                     1.316e-2;1.00000
        (NT2692 NT2239 NT2086 NT2084 NT2087 NT2082) 0.171 ;4.00000
        (NT2692 NT2709 NT2082)                      0.434    ;9.00000
        (NTDATE-7896-2081 NT2082 NT2083 NT2084 NT2109 NT2709) 0.382 ;8.0
NT2082      address                                  1.000    ;22.0000
NT2083      dear                                     1.000    ;22.0000
NT2084      person                                   1.000    ;43.0000
NT2086      sincerely                                1.000    ;17.0000
NT2087      jobtitle                                 1.000    ;13.0000
NT2109      best                                     3.899e-2;5.00000
            paragraph                                0.883    ;97.0000
            regards                                  3.899e-2;5.00000
            with                                     3.899e-2;5.00000
NT2121      devotedly                                0.375    ;2.00000
            warmly                                   0.625    ;3.00000
NTDATE-7896-2081    date                             1.000    ;22.0000
```

Figure 9. Induced SCFG for the Correspondence Record Type.

Nonterminals are of the form S or NTnnnn. Nonterminals on the left-hand side of Figure 9 are the left-hand side of rules. The nonterminal S is the initial symbol. Nonterminals in parentheses in the second column are the corresponding right-hand-side of rules. The probability of a rule is shown in the third column. The last column shows the number of times the corresponding rule is used to generate one of the strings in the sample.

The grammatical induction method started with twenty-one rules with S on the left-hand side and the intellectual elements from the sample on the right-hand sides. It then added twelve rules with a nonterminal on the left-hand side and each distinct intellectual element on the right. Applying the chunking and merge operators, it managed to reduce the number of rules needed to describe the structure of the sample to twelve rules. The initial symbol now describes four general structures from which the entire sample can be derived.

Figure 10 shows a grammar created by the authors of this paper. In our opinion, it is a better grammar for this sample. In addition to the twelve rules that have a nonterminal generating an intellectual element, there are only eight other rules. The initial symbol S describes two general forms of correspondence. It also captures the recursive structure of the BODY of correspondence and that "With best regards" can only appear at the end of the BODY.

S → DATE ADDRESS SALUTATION BODY CLOSE
S → DATE SALUTATION BODY CLOSE ADDRESS
SALUTATION → DEAR PERSON
BODY → PARA BODY
BODY → PARA
BODY → W B R
CLOSE → FOR PERSON
CLOSE → FOR PERSON JOBTITLE
DATE → date
DEAR → Dear
ADDRESS → address
PERSON → person
JOBTITLE → jobtitle
PARA → paragraph
FOR → Sincerely
FOR→ Devotedly
FOR → Warmly
W → With
B → best
R → regards

Figure 10. A Better Grammar for White House Correspondence.

With a larger sample, the grammatical induction method could be expected to converge to grammar more comparable to that generated by the authors. Of course, the larger sample would undoubtedly have additional elements and configurations of elements that would complicate the grammar. That is one of the major reasons for desiring to infer this grammar automatically. It is too time consuming to try to capture the variability in style of hundreds of different writers of correspondence.

3.3.2 A Grammar for the Documentary Form of White House Memoranda

The grammatical induction method was applied to a sample of twenty-six White House memoranda (See Appendix A). Figure 11 shows an e-record in the sample. [3]

[3] Bush Presidential Library, Bush Presidential Records, Council of Economic Advisors, Michael J. Boskin's Files

May 4, 1989

MEMORANDUM FOR THE PRESIDENT

FROM: MICHAEL J. BOSKIN

SUBJECT: Employment and Unemployment in April, Labor Department Release, Tomorrow Morning, 8:30 a.m

According to the household survey. the unemployment rate for civilian workers rose 0.3 percentage point to 5.3 percent in April. The unemployment rate for all workers, including the armed forces, also rose 0.3 percentage point to 5.2 percent. The unemployment rates for teenagers and Hispanics recorded the largest increases. Nevertheless, the employment to population ratios for all workers and for civilian workers remained at their all-time highs of 63.3 percent and 63.0 percent, respectively.

Nonfarm payroll employment rose modestly in April. Nonfarm employment increased by 117,000 jobs in April according to the survey of business establishments. The April increase was less than the revised March increase of 171,000 jobs, which was depressed by the strike at Eastern. The payroll employment gains in April were the smallest since June 1986.

According to the payroll survey, employment in the goods-producing sector rose by 5,000 jobs, while employment in the service-producing sector rose by 112,000 jobs. Manufacturing jobs fell by 9,000 after rising a revised 6,000 jobs in March. Average weekly hours in manufacturing rose by 0.3 hours to 41.3 and manufacturing overtime hours rose by 0.1 hour to 4.0 hours, but these gains may reflect seasonal adjustment problems. Average hourly earnings in the nonfarm sector rose by 0.7 percent to $9.59 from a revised $9.52 in March.

Figure 11. Sample White House Memorandum.

The information extraction algorithm applied to this record produced the XML annotated document shown in Figure 12.

```
                                    <paragraph><Date>May 4, 1989</Date>
</paragraph>

<paragraph>MEMORANDUM FOR <Person>THE PRESIDENT</Person>
</paragraph>
<paragraph>FROM:              <Person>MICHAEL J. BOSKIN</Person>
</paragraph>
<paragraph>SUBJECT:          Employment and Unemployment in <Date>April</Date>, <Organization>Labor
                        Department</Organization> Release, <Date>Tomorrow Morning</Date>, <Date>8:30 a.m</Date>
</paragraph>
          <paragraph>According to the household survey. the unemployment rate for
civilian workers rose 0.3 percentaqe point to <Percent>5.3 percent</Percent> in
<Date>April</Date>. The unemployment rate for all workers, including the
armed forces, also rose <Percent>0.3 percentage</Percent> point to <Percent>5.2 percent</Percent>.
The unemployment rates for teenagers and Hispanics recorded the
largest increases. Nevertheless, the employment to population
ratios for all workers and for civilian workers remained at their
all-time highs of <Percent>63.3 percent</Percent> and <Percent>63.0 percent</Percent>, respectively.
</paragraph>
          <paragraph>Nonfarm payroll employment rose modestly in <Date>April</Date>. Nonfarm
employment increased by 117,000 jobs in <Date>April</Date> according to the
survey of business establishments. The <Date>April</Date> increase was less
than the revised <Date>March</Date> increase of 171,000 jobs, which was
depressed by the strike at Eastern. The payroll employment gains
in <Date>April</Date> were the smallest since <Date>June 1986</Date>.
</paragraph>
          <paragraph>According to the payroll survey, employment in the goods-
producing sector rose by 5,000 jobs, while employment in the
service-producing sector rose by 112,000 jobs. Manufacturing jobs
fell by 9,000 after rising a revised 6,000 jobs in <Date>March</Date>.
Average weekly hours in manufacturing rose by <Date>0.3 hours</Date> to 41.3
and manufacturing overtime hours rose by <Date>0.1 hour</Date> to <Date>4.0 hours</Date>,
but these gains may reflect seasonal adjustment problems.
Average hourly earnings in the nonfarm sector rose by <Percent>0.7 percent</Percent>
to <Money>$9.59</Money> from a revised <Money>$9.52</Money> in <Date>March</Date>.
</paragraph>
```

Figure 12. XML Markup of Named Entities in the Sample Memo

This document is not an XML document, but includes XML tags annotating semantic categories, such as dates and person's names. The content removal algorithm is applied to the marked-up records leaving a list of the intellectual elements of documentary form of each record in the sample. Figure 13 shows the intellectual elements of the documentary form of the twenty-six memoranda. The first list of intellectual elements in the figure corresponds to the memorandum in Figs. 11 and 12.

```
(date MEMORANDUM FOR person FROM person SUBJECT np paragraph paragraph paragraph)
;;026.txt
(date MEMORANDUM FOR person FROM person SUBJECT np paragraph paragraph) ;;078.txt
(date MEMORANDUM FOR person FROM person SUBJECT np paragraph heading paragraph paragraph
heading paragraph heading paragraph paragraph paragraph paragraph paragraph heading
paragraph paragraph paragraph paragraph heading paragraph paragraph paragraph heading
paragraph paragraph paragraph) ;;094
(date MEMORANDUM FOR person FROM person SUBJECT np paragraph paragraph paragraph)
;;095.txt
(date MEMORANDUM FOR person FROM person SUBJECT np paragraph) ;;109.txt
(date MEMORANDUM FOR person person FROM person SUBJECT np heading paragraph heading
heading paragraph paragraph paragraph heading paragraph heading paragraph paragraph
heading paragraph paragraph paragraph paragraph paragraph) ;;128.txt
(date MEMORANDUM FOR person FROM person SUBJECT np heading paragraph paragraph
headingparagraph paragraph paragraph heading paragraph paragraph paragraph heading
paragraph paragraph paragraph paragraph heading paragraph paragraph paragraph paragraph
paragraph) ;; 129.txt
(date MEMORANDUM FOR organization FROM person person SUBJECT np paragraph paragraph) ;;
130.txt
(date MEMORANDUM FOR person THROUGH person jobtitle FROM person jobtitle SUBJECT np
paragraph paragraph paragraph paragraph) ;;131.txt
(date MEMORANDUM FOR person FROM person person SUBJECT np heading paragraph heading
paragraph paragraph paragraph heading paragraph paragraph paragraph heading paragraph
paragraph paragraph paragraph heading paragraph paragraph paragraph paragraph paragraph
paragraph paragraph paragraph paragraph paragraph paragraph paragraph) ;; 132.txt
(date MEMORANDUM FOR person jobtitle FROM person jobtitle SUBJECT np paragraph paragraph
paragraph paragraph paragraph) ;; 133.txt
(date MEMORANDUM FOR person FROM person jobtitle SUBJECT np paragraph paragraph paragraph
paragraph paragraph) ;;144.txt
(date MEMORANDUM FOR person THROUGH person FROM person SUBJECT np paragraph paragraph
paragraph paragraph paragraph paragraph) ;;143.txt
(date MEMORANDUM FOR person FROM person SUBJECT np paragraph) ;; 032
(date MEMORANDUM FOR person jobtitle FROM person jobtitle SUBJECT np paragraph paragraph
paragraph paragraph paragraph) ;; 134
(date MEMORANDUM FOR person FROM person SUBJECT np paragraph paragraph paragraph
paragraph paragraph paragraph) ;; 137
(date MEMORANDUM FOR person FROM person SUBJECT np heading paragraph) ;;049a
(date MEMORANDUM FOR person FROM person SUBJECT np heading paragraph paragraph heading
paragraph) ;;142.txt
(date MEMORANDUM FOR organization FROM person SUBJECT np heading paragraph paragraph
paragraph paragraph) ;;025.txt
(date MEMORANDUM FOR person FROM person jobtitle SUBJECT np heading paragraph) ;;049b
(date MEMORANDUM FOR organization FROM person person SUBJECT np paragraph heading
paragraph) ;;093.TXT
(date MEMORANDUM FOR organization FROM person jobtitle SUBJECT np paragraph paragraph)
;;091.txt
(date MEMORANDUM FOR organization FROM person SUBJECT np paragraph paragraph paragraph
paragraph) ;;0141.TXT
(date MEMORANDUM FOR organization FROM person SUBJECT np paragraph paragraph) ;;146
(date MEMORANDUM FOR person FROM person SUBJECT np heading paragraph paragraph paragraph
paragraph paragraph paragraph) ;;102.txt
(date MEMORANDUM FOR organization FROM person SUBJECT np paragraph paragraph paragraph
paragraph paragraph paragraph) ;;040.txt
```

Figure 13. Intellectual Forms of Sample White House Memoranda

The grammar induced from this small sample of memoranda is shown in Figure 14.

```
S ->  A H I E M
      A G
      A I F
      B M
      B
      C G
A ->  DATE MEMORANDUM FOR J
      A THROUGH H
B ->  C J D
      C D M
      B F
      B E
C ->  A K
D ->  J L L
E ->  F F
      E E E
F ->  M M
G ->  G M
      I E
H ->  J K
I ->  K H L L
      D B
J ->  PERSON
      ORGANIZATION
K ->  FROM
      JOBTITLE
L ->  SUBJECT
      NP
M ->  PARAGRAPH
      HEADING
```

Figure 14. A Context-Free Grammar Induced from Sample White House Memoranda

To make the grammar easier to understand, the nonterminal symbols of the form NTnnnn have been replaced by capital letters. Furthermore, the twelve rules with a nonterminal on the left-hand side and the terminal (intellectual) elements on the right-hand side (person, organization) are not shown. The grammatical induction method began with twenty-six rules with the start symbol S on the left-hand side and the sequence of intellectual elements on the right-hand sides. It has produced a grammar with thirty rules with fourteen nonterminals (S, A-M), excluding the nonterminals PERSON, ORGANIZATION, etc. It has inferred that memos are of six primary forms, as indicated by the fact that there are six rules that begin the start symbol S. It has recognized that all of the memoranda begin with DATE MEMORANDUM FOR J, where J is a PERSON or ORGANIZATION. It has not yet induced a recursive rule for the body of a memo made up of a sequence of PARAGRAPHs or sections that begin with a HEADING followed by a sequence of PARAGRAPHs.

Figure 15 shows a grammar for White House Memoranda created by the authors of this paper. We have left out the twelve rules of the form "PERSON -> person". There are seventeen rules and nine nonterminals.

MEMO → HEAD BODY
HEAD → PREFIX FROM ENTITIES SUBJECT NP
PREFIX → DATE MEMORANDUM FOR ENTITIES
PREFIX → PREFIX THROUGH ENTITY
BODY → PARAGRAPHS
BODY → SECTIONS
BODY → PARAGRAPHS SECTIONS
PARAGRAPHS → PARAGRAPH PARAGRAPHS
PARAGRAPHS → PARAGRAPH
SECTIONS → SECTION SECTIONS
SECTIONS → SECTION
SECTION → HEADING PARAGRAPHS
ENTITIES → ENTITY ENTITIES
ENTITIES → ENTITY
ENTITY→ PERSON
ENTITY → PERSON JOBTITLE
ENTITY → ORGANIZATION

Figure 15. A Better Grammar for White House Memoranda

The structural description produced by this grammar is more succinct and intuitive than the one produced by the grammatical induction method. It is to be expected that with a larger sample of memoranda with additional intellectual elements, the grammatical induction method will converge to a more concise grammar.

4. Method for Recognizing Documentary Form

Having induced or inferred grammars for a variety of record types, how can they be used to recognize the documentary structure of personal computer records and thus identify record types? Our approach to recognizing documentary structure is illustrated in Figure 16.

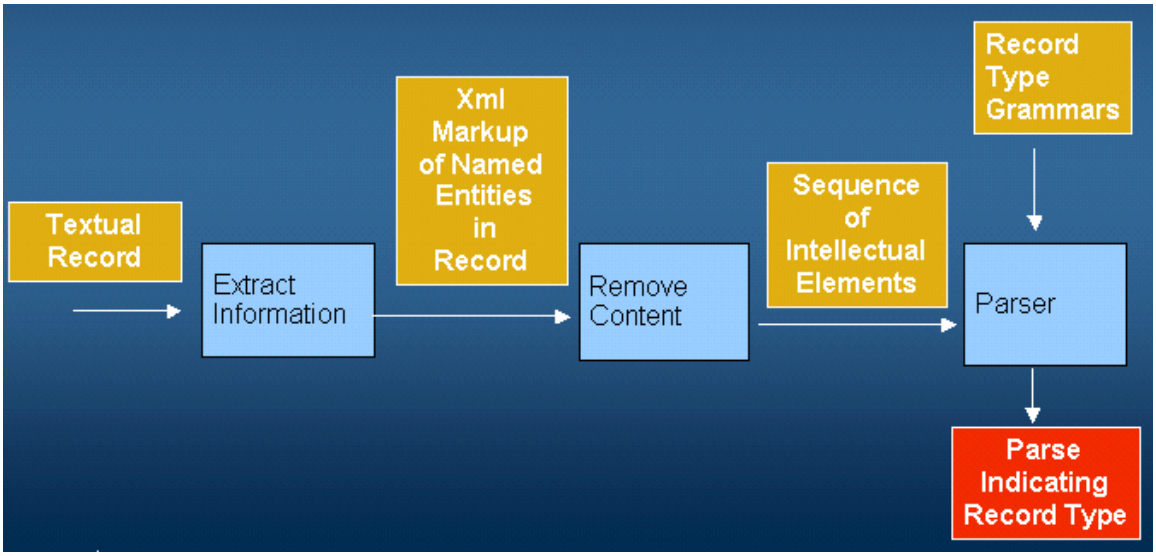


Figure 16. Method for Recognizing Documentary Form.

A textual record of unknown record type is converted to a text or html file format. The information extractor identifies named entities in the record and uses XML markup to indicate these entities. The content removal program produces a list of intellectual elements of the record. Finally, a context-free grammar parser uses the induced grammars to parse the list of intellectual elements. The parse tree describes the documentary form and indicates the record type.

Shown below is a list of intellectual elements derived from a record using the above method.

(date address dear person paragraph paragraph sincerely person jobtitle)

Using the grammars induced for correspondence and memoranda, the parser for stochastic context-free grammars produces the following parse tree nested parenthesis notation.

(S (ntdate-7896-2081 . date) (nt2082 . address) (nt2083 . dear)
(nt2084 . person) (nt2109 . paragraph)
(nt2709
(nt2709 (nt2109 . paragraph) (nt2086 . sincerely) (nt2084 . person)))
(nt2087 . jobtitle)))

Figure 17. Parse Showing the Documentary Form of a Record.

The grammar that produced this parse is the grammar inferred for correspondence. Therefore, the record type is correspondence.

Figure 18 shows the same parse as a graphical tree.

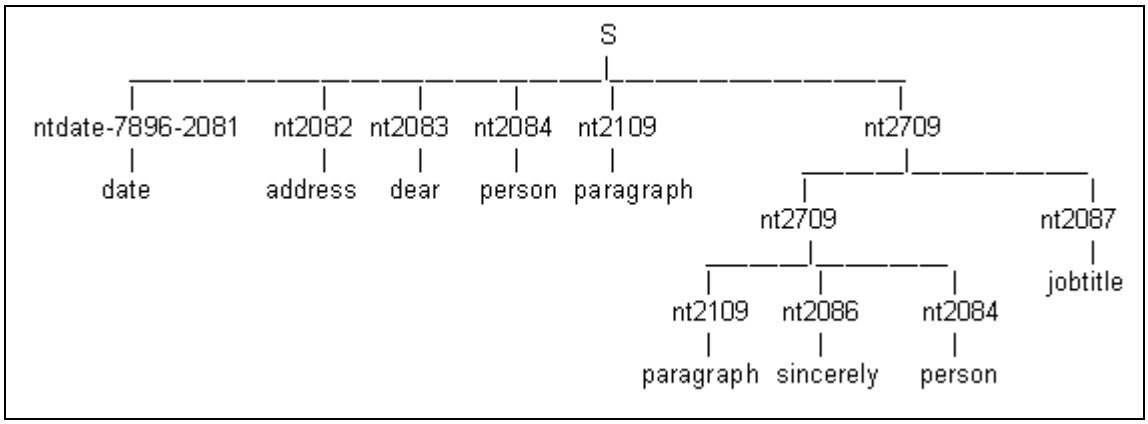


Figure 18. Graphical Parse Tree Showing Documentary Form of a Record

This parse tree also shows, as noted before, that the induced grammar does not yet include recursive rules that represent the body of correspondence as a sequence of paragraphs. This is due to the small sample size. In addition, there is not a rule that associates the person's name following the formula of respect, sincerely, with the jobtitle. This is due in part to the grammatical induction method only allowing a sequence of elements as input. The information extraction algorithm, on the other hand, uses <paragraph> tags to indicate that "person jobtitle" and "dear person" are contiguous. In future research, the grammatical induction method will be extended to allow sequences of elements and contiguous elements. This may also speed up the generalization of recursive rules for paragraphs since there may be less likelihood of their being associated with elements that are contiguous.

5. Related Research

Researchers concerned with classifying images of paper documents have also studied document structure by their differences in layout. Document structures can be constructed by bottom-up analysis of graphical layout instead of textual content. The quantity and location of white space in a block of text is analyzed in order to decompose it into a hierarchical set of divisions, such as lines, paragraphs, and sections [Rus and Summers, 1997]. Document type classification can be accomplished without OCR by introducing an interval encoding that captures elements of the spatial layout of the document and then classifying of the documents using Hidden Markov Models (HMMs) [Hu et al. 1999a, 1999b]. In the WISDOM++ system, decision trees and first-order logic learning is used to classify segments or blocks of text and inductively learn rules for layout-based classification [Esposito et al 2000].

Another approach to classifying documents uses ASCII text [Ma et al 2003]. The system they create uses structural synopses that contain the most important features. A decision tree over these features is used for classification.

The Sequitur algorithm [Nevill-Manning and Witten 1997] takes as input a single sequence and returns a hierarchical decomposition of it. The algorithm looks only for exact repetitions and is incapable of generalizing to recursive or disjunctive rules.

Keller and Lutz's algorithm applies a genetic algorithm to infer a SCFG given a set of positive examples [Keller and Lutz 1997]. The algorithm tries to optimize the parameters for a grammar given a set of language examples. The fitness function used is based on minimum description length (MDL) principle that biases the search to favor models that describe the data more simply while taking into consideration the simplicity of the model given the data.

Heeringa and Oates describe two different algorithms (SPAN and PRESPAN) for learning the parameters of SCFGs [Heeringa and Oats 2001]. Both of these algorithms compute the rule probabilities or parameters of an SCFG given its structure and a set of sentences that are generated by the grammar. Both algorithms then use a chart parser to parse each sentence and keep a histogram that tracks the number of times a rule is used when parsing an individual sentence. SPAN and PRESPAN are both online algorithms that execute in a fixed amount of space regardless of the number of sentence observations and perform comparably to the inside-outside algorithm for parameter estimation. The two algorithms differ in "the selection of rules to update, criteria for updates, and the update rule itself."

Klein and Manning claim that one of the drawbacks with MDL or Bayesian approaches using expectation maximization (EM) for grammar induction is that they fail in producing parse trees that are more in line with linguistically-determined grammars. Furthermore, they claim that these approaches fail "to propagate contextual cues efficiently." Their Constituent Context Model makes two assumptions about constituents. First, constituents of a parse do not cross. Secondly, constituents occur in constituent contexts. Their model only represents the conditional likelihood of the data (trees) and an EM-like algorithm is used to find the model that locally maximizes the conditional likelihood of the data [Klein and Manning 2002].

Hong [2003] describes an algorithm that does a hill-climbing search through the space of possible stochastic context-free grammars. The initial grammar is a set of rules in which the left-hand side is the initial symbol and the right-hand sides are the strings in the language sample. Given a set of transformation rules, all possible one-step transforms are generated. The complexity of each grammar is evaluated. The complexity function for evaluating grammars is expressed in terms of rule lengths, number of unique rules, and number of unique symbols in rules. The grammar with the lowest complexity is selected. This process is iterated until all next transformations of the grammar have higher complexities than the current grammar. There is no guarantee that the given set of operators generates all possible SCFGs for the given sample. However, that may not be necessary as an intuitive, "linguistically-determined grammar" may be best.

The Bayesian approach investigated in this paper [Stolcke and Omohundro 1994] and the minimum description length (MDL) approach [Keller and Lutz 1997] are very similar. The major difference between these two algorithms is their searching method and

operators. The Bayesian model merging approach for inducing stochastic context-free grammars was used because the search strategy provides a local optimum whereas genetic algorithms do not guarantee this. The advantage that Bayesian model merging approach has over SPAN and PRESPAN is that it uses a form of the inside-outside algorithm that SPAN and PRESPAN approximate, while also providing operators to construct the grammar, whereas SPAN and PRESPAN assume the grammar plus examples. Klein and Manning's observations coincide with the author's regarding the importance of considering constituent structure and context in performing grammatical induction.

6. Results and Research Issues

Methods have been developed and demonstrated for:

- Annotating many of the semantic categories that comprise the intellectual form of a document.
- Removing the content of an annotated document and leaving the intellectual form.
- Inducing (inferring) a grammar for the documentary structure of a record type from a sample of intellectual elements of records of the same type.
- Recognizing the documentary form of a record by parsing its intellectual elements using the inferred grammars for different record types.

This research has used small samples of just two record types, correspondence, and memoranda. Experiments need to be conducted using larger samples of records from the Bush personal computer staff files. Furthermore, experiments are needed to learn a larger variety of record types, for example, press releases, resumes, decision memoranda, Executive Orders, schedules, mailing lists, and transcripts of news conferences. Then experiments should be conducted to evaluate the performance of the induced grammars in identifying the record types of Presidential e-records.

There are additional intellectual elements that need to be recognized in textual e-records, for example, page numbers, section headings, courtesy copies, lists, and attachments. Some of these elements are dependent on physical elements of form such as font, boldface, italics, capital letters, bullets, and numbering. Demonstrations in this paper were performed on an ASCII text version of the original document. Experiments should be conducted using HTML versions of the original text document to learn elements of physical form as well as intellectual form.

The grammars that are induced from a sample of documents of a particular type often do not describe the documentary form the same way that a person would describe its form. This seems to be in part because it does not allow as input indications of elements that are more closely associated that others. Modifications will be made to the method to enable the inclusion of this kind of information in the input strings. This may also be the result of the induction method not accounting for the semantics of the input string. The use of a reference grammar that contains some rules that should be included in inducing the

grammar will be explored. This will bias the induced grammar to include some conventional terms and units for the structure.

The current technique for inducing grammars describing documentary form uses only two operators, chunking and merging. These two operators are capable of producing all possible grammars for a sample, and even generalizing to include strings not in the sample. However, the alternative grammars generated and the evaluation function used to select grammars produce some grammars that describe structures that are non-intuitive, that is, a person would not describe the structure of the document in that way. The kinds of operators should be extended to see whether grammars that are more intuitive could be induced. Hong's set of transformations may provide some advantages over the two transformations used in Stolke's model merging procedure.

References

[Duranti 1998] L. Duranti. *Diplomatics: New Uses for an Old Science* (Lanham, Md.: Society of American Archivists and Association of Canadian Archivists in association with Scarecrow Press, 1998.

[Esposito et al 2000] F. Esposito, D. Malerba and F. A. Lisi, Machine Learning for Intelligent Processing of Printed Documents, *Journal of Intelligent Information Systems*, Volume 14, No. 2-3, pp.175-198, 2000.

[Heeringa and Oats 2001] B. Heeringa and T. Oates. Two Algorithms for Learning the Parameters of Stochastic Context-Free Grammars. Using Uncertainty Within Computation. Papers from the *AAAI Fall Symposium*. November 2-4, 2001, North Falmouth, Massachusetts, Technical Report FS-01-04.

[Hong 2003] T. W. Hong, Grammatical Inference for Information Extraction and Visualisation on the Web. PhD Dissertation, Department of Computing, Imperial College of Science, Technology, and Medicine, London, UK, June 2003.

[Hu et al 1999a] J. Hu, R. Kashi, and G. Wilfong**.** Document Classification using Layout Analysis. *Proceedings of the Workshop on Document Analysis and Understanding for Document Databases (DAUDD'99),* pp. 556-560, Florence, Italy, September 1999

[Hu et al 1999b] J. Hu, R. Kashi and G. Wilfong. Document Image Layout Comparison and Classification. *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'99),* pp. 285-288, Bangalore, India, September 1999.

[InterPARES 2001a] "Authenticity Task Force Report," *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project.* www.interpares.org/book/interpares_book_d_part1.pdf

[InterPARES 2001b] *The InterPARES Glossary: A Controlled Vocabulary of Terms Used in the InterPARES Project* No. 2, Vol. 1 (Vancouver: University of British Columbia, 2002) www.interpares.org/documents/InterPARES%20Glossary%202002-1.pdf

[Isbel and Underwood 2006] S. Isbell and M. Underwood. The PERPOS Information Extractor Applied to Presidential E-Records. Working Paper 05-10, February 2006.

[ISO 1986] International Standards Organization. Standard Generalized Markup Language - ISO 8879.

[Iwanska and Underwood 2006] L. Iwanska and W. E. Underwood. Natural Language Boolean Queries. Working Paper 06-01, February 2006

[Keller and Lutz 1997] B. Keller and R. Lutz. Evolving stochastic context-free grammars from examples using a minimum description length principle. In *Workshop on Automata Induction, Grammatical Inference and Language Acquisition*. ICML97, 1997.

[Klein and Manning 2002] D. Klein and C. D. Manning. Natural language grammar induction using a constituent-context model. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems* 14, Cambridge, MA, 2002. MIT Press.

[Ma et al 2003] L. Ma, J. Shepherd, and A. Nguyen, Document classification via structure synopses, in *Proceedings of the Fourteenth Australasian Database Conference on Database technologies* 2003.

[Nevill-Manning and Witten 1997] C. G. Nevill-Manning and Ian H. Witten. Compression and explanation using hierarchical grammars. *The Computer Journal*, Vol 30 Issue 3, 1997.

[Pearce-Moses 2004] R. Pearce-Moses. *A Glossary of Archival and Records Terminology*, Society of American Archivists, 2004. www.archivists.org/glossary/index.asp

[Rus and Summers 1997] D. Rus and K. Summers. Geometric algorithms and experiments for automated document structuring. *Mathematical and Computer Modelling*, 1997. 26(1):55-83.

[Stolcke and Omohundro 1994] A. Stolcke and S. Omohundro. Inducing Probabilistic Grammars by Bayesian Model Merging. *Lecture Notes In Computer Science*; Vol. 862**,** *Proceedings of the Second International Colloquium on Grammatical Inference and Applications* 1994 , pp.106 - 118.

[Underwood 2005] W. E. Underwood. Automatic Description of the Content of Presidential Record Series, Working Paper 05-09, GTRI, July 2006.

[Underwood and Hayslett-Keck 2004] W. Underwood, M. Hayslett-Keck. A Corpus of Presidential, Federal and Personal Records for use in Information Extraction, Description and FOIA/PRA Review Experiments, PERPOS Technical Report ITTL/CSITD 04-5, June 2004.

[Underwood and Harris 2005] W. E. Underwood and B. Harris. The Knowledge and Reasoning Required to Recognize Presidential Record Act Restrictions and Personal Record Misfiles. PERPOS Working Paper ITTL/CSITD 05-03, Georgia Tech Research Institute, 2005.

[W3C 1999a] World Wide Web Consortium. HTML 4.01 Specification, 24 Dec 1999. www.w3.org/TR/html4/sgml/dtd.html

[W3C 1999b] World Wide Web Consortium. XSL Transformations (XSLT) Version 1.0, 16 Nov 1999. www.w3.org/TR/xslt

[W3C 2000] World Wide Web Consortium. XHTML™ 1.0: The Extensible HyperText Markup Language, 26 Jan 2000.

[W3C 2001] World Wide Web Consortium. Extensible Stylesheet Language (XSL) Version 1.0, 15 Oct 2001

[W3C 2006a] World Wide Web Consortium. Extensible Markup Language XML 1.0 (Fourth Edition) 16 Aug 2006. www.w3.org/TR/REC-xml/

[W3C 2006b] World Wide Web Consortium. Cascading Style Sheets level 1, Revision 1, CSS 2.1 Specification 11 April 2006.

Appendix A: Sample Correspondence and Memoranda Record Types

The sample documents are described in greater detail by Underwood and Hayslett-Keck [2004].

White House Outgoing Correspondence

Document No.	Description
001	A letter from Barbara Bush
005	A letter from President Bush to his granddaughter
006	A letter from President Bush to his son George W. about PFC James Markwell
007	Letter From President Bush to Head of State Deng Xiaoping
008	Letter from President Bush to Mikhail Gorbachev
009	Letter President Bush to Republican Representative Jim Lightfoot
016	Letter from President Bush to Lee Atwater
029	Letter from Nicholas Calio to Wayne Gilchrest
031	Letter to the President from C. Boyden Gray
037	Letter to Phillip Lochner from C. Boyden Gray
059	Letter to Mr. Augustine from James P. Pinkerton, Deputy Assistant to the President for Policy Planning
068	Letter to Sister Katherine T. McNamee from John Sununu
070	Letter to Ambassador Melady from Jane Barnett Leonard, Assistant Director Office of Public liaison
072	Nicholas E. Calio , Assistant to the President for Legislative Affairs to Senator Pressler
076	Letter from Nicolas Calio to Representative Anthony
080	Letter to Richard Swan from Jeffrey W. Vogt. Assistant Director Office of the Public Liaison
082	Letter to Peter Ruane from John H. Sununu , Chief of Staff
084	Letter to Ray Allen from Doug Wead, Special Assistant to the President for Public Liaison
086	Letter to Lucille Anderson from Doug Wead, Special Assistant to the President for Public Liaison
106	Letter from Barbara Bush to Marge Simpson
107	Letter to Fred Kleinknecht from Barbara Bush
112	Letter to Mrs. Liard from Shirley Green

White House Internal Memoranda

Document No.	Description
025	Memo to the President from Mike Boskin
026	Memo to the President from Mike Boskin
032	Memo for the President from Constance Horner
040	Memo for the President from Roger B. Porter
049	Memo for the President from Chase Untermeyer
078	Memo for Sam Skinner from Ede Holiday
091	Memo for John Sununu from Lester South
092	Mmemo to Chase Untermeyer from Fred French
093	Memo to John Sununu from Roger Porter
094	Memo to President from Roger Porter
095	Memo to Boyden Gray from Nelson Lund
102	Memo from Jim Cicconi to the Chief of Staff concerning a House Resolution
109	Memo for the President
128	Memo to Fred McClure and Roger Porter from Nicholas Calio concering Minimum Wage
129	Memo from Anne Brooks Gwaltney to Chase Untermeyer concerning National Endowment for the Arts
130	Memo to the National Advisory Council on International Monetary and Financial Policies from John Robson concerning Export Credits to Iraq
131	Memo for Michael Baroody from Zelda Novak concerning Minimum Wage
132	Memo for John Sununu from Allan Bromley and David Bates concerning Global Climate Change Convention Negotiations
133	Memo from John Niehuss for Stephen Danzansky concerning World Bank Green Fund
134	Memo for Nicholas Brady from Gregg Persmeyer concerning Charitable Deductions
137	Memo for Samuel Skinner from Roger Porter concerning letter to Magic Johnson
141	Memo for Doug Wead from Les Csorba concerning Tower Nomination
142	Memo for Doug Wead from Shirly Green concerning Space Planning Issues
143	Memo for Marlin Fitzwater from Clayton Fong concerning President's news Conference
144	Memo for Governor Sununu from Jeff Vogt concerning recommended phone call
146	Memo for David Demarest form Joe Watkins concerning Bush Strategy for Black America