DRAFT

# Semantic Annotation of Presidential E-Records

Sheila Isbell
Matthew Underwood
William Underwood

Technical Report ITTL/CSITD 07-01

August 2007

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 W911NF-06-2-0050. The findings in this paper should not be construed as an official ARL or NARA position unless so indicated by other authorized documentation.

## Abstract

Archivists could gain earlier intellectual control of large volumes of accessioned electronic records if it were possible to automatically describe items, file units and record series. Archivists could be more productive in reviewing Presidential e-records if metadata could be automatically extracted to fill in information needed in withdrawal forms (e.g. document type, chronological date, author(s), addressee(s), and subject). A capability is needed to perform searches for e-records relevant to FOIA requests that allows search on document type, author's names, addressee's names, chronological date and topics.

Each of these opportunities to improve archival processing of electronic records depends on the capability to automatically recognize and annotate semantic categories in text such as person's names, dates, job titles, and postal addresses. The capability to recognize document types such as correspondence, memoranda, schedules, minutes of meetings and press releases also depends on the capability to automatically recognize semantic categories in text.

The results of a previous information extraction experiment were analyzed to determine reasons for partially correct, missing annotations and false positives. The solution to the problems were in the provision of additional wordlists for primitive semantic categories and in additional or modified JAPE rules.

Additional wordlists and JAPE rules were created for recognizing and annotating additional semantic categories, namely relative temporal expressions, and congressional bills and statutes. Additional JAPE rules were constructed for disambiguating terms that might designate either locations or facilities, and disambiguating terms that might designate countries as geographic areas, governments of the country, or citizens of the country.

Previously, experiments were conducted on copies of paper Presidential records that were scanned and OCRed. The current experiments will be conducted using e-records from the Bush Presidential personal computer records. Two experiments will be conducted. The first experiment will evaluate the performance of the information extractor with regard to the named entities addressed in the previous two experiments, namely, annotation of person, location, organization, date, money and percent. The second experiment will address the annotation of additional semantic categories, namely job titles, social security numbers, addresses, facilities, congressional bills and statutes, relative temporal expressions and terms that might be the names of countries, governments or citizens..

# Table of Contents

# 1. Introduction

## *1.1 Background*

Archivists could gain earlier intellectual control of large volumes of accessioned electronic records if it were possible to automatically describe items, file units and record series. Archivists could be more productive in reviewing Presidential e-records if metadata could be automatically extracted to fill in information needed in withdrawal forms (e.g. document type, chronological date, author(s), addressee(s), and subject). A capability is needed to perform searches for e-records relevant to FOIA requests that allows search on document type, author's names, addressee's names, chronological date and topics.

Each of these opportunities to improve archival processing of electronic records depends on the capability to automatically recognize and annotate semantic categories in text such as person's names, dates, job titles, and postal addresses. The capability to recognize document types such as correspondence, memoranda, schedules, minutes of meetings and press releases also depends on the capability to automatically recognize semantic categories in text.

In prior research, the capability to recognize and annotate person’s names, location names, dates and organization names in textual Presidential e-records was developed and demonstrated. The capability to recognize job titles and mailing addresses was also developed. The capability of the PERPOS information extractor to recognize person's names, location names, organization names, job titles, dates and mailing addresses in Presidential e-records was experimentally evaluated [Isbell et al 2006].

This technology is based on wordlists containing the names of entities in semantic categories such as person's first and last names, job titles, organization names and city and state names, and on processing resources that annotate text based on the terms that appear in these wordlists. It is also based on rules that apply to annotated text to produce more complex annotations such as person's full names, mailing addresses, relative temporal expressions, and on rules to discriminate organization names from facilities and country names from government names.

## *1.2 Purpose*

During the current year of research, the results of the previous experiment were analyzed to identify additions to the wordlists and changes to the rules that were needed to improve the performance of the semantic category annotator. Extensions to the word lists and rules were made to recognize additional semantic categories such social security numbers and congressional bills and acts. The purpose of this report is to describe these additions to the wordlists, changes to the rules and experiments to evaluate the performance of the enhancements of the semantic annotation system.

### *1.3 Scope*

In section 2, extensions to the wordlists are discussed. In section 3, improvements to the Java Annotation Pattern Engine (JAPE) rules are discussed. In section 4, tests to verify improvements are described. In section 5, scalability of the semantic annotation method is discussed. In section 5, experiments in annotating Bush personal computer records are discussed. In section 7, results are summarized.

## 2. Improvements to Wordlists

This section describes extensions to wordlists for person's names, titles and position names, location and facility names, organization names, temporal expressions, money terms and titles of congressional bills and statutes.

### *2.1 Person's Names*

In a previous experiment [Isbell et al 2006], the experimental corpus contained 463 occurrences of person's names. Of these, 333 person names were annotated correctly; 66 were partially correct annotations; 64 person names were not annotated and there were 145 false positives

The list of international female first (given) names provided with ANNIE contains 5818 names. The US Census Bureau has 4275 most frequent female first names (occurring in the 1990 census.[1] These names occurred in 90% of the US population. These two lists were merged producing a list of 8380 female first names.

In the previous experiment, two lists of male first names were used. The one provided with ANNIE contains 4373 names. The second list consists of the 1216 most common male names in the 1990 Census. These names occurred in 90% of the US population in 1990. These two lists were merged to create a list of male first names with 4788 entries.

The US Census Bureau also created a list of 88,799 of the most frequent surnames in the 1990 US Census. These names occurred in 90% of the US population. This list replaces the person_lastTopUS.lst which contained the 2500 most frequent last names in the 1990 US Census. It is now named surname_US_90_census.lst.

The following lists should be merged, compared with the surnames_US_90_census list, and a list of US-surnames_supplement created.

| | |
|---|---|
| Appointees-1989Lastames.lst | 60 |
| LastNames_Representative_in-101_102_congress.lst | 451 |
| Nominee_last_names.lst | 364 |
| Senators_last_names_101-102_congress.lst | 107 |
| Whitehousestafflastnames.lst | 395 |

[1] http://www.census.gov/genealogy/names/names_files.html

The names of previous US Presidents are often mentioned in Presidential records. A wordlist containing the names of the 43 presidents plus some of their nicknames has been created.

Presidential records discussing legislative strategy often mention congressmen by name. The full names of senators and representatives of the 101st and 102nd Congresses have been extracted from the Congressional Record [LOC 1990, 1992] and used to create word lists.

The names of foreign chiefs of state and heads of government often appear in transcripts of press conferences, press releases, notes of meetings, and correspondence. A list of the full names of heads of state and heads of governments was created from the 1990 CIA World Factbook [CIA 1990]. The names of Ambassadors to the United States also appear in the CIA World Factbook. It would be useful to have these as well, but they have not yet been extracted.

Figure 1 summarizes the additions and modifications made to lists of person's names.

| **Wordlist** | **Description** | **Major Type** | **Minor Type** | **Count** |
|---|---|---|---|---|
| person_female_first.lst | Female first names from US 1990 census merged with ANNIE's default list | person_first | female | 8380 |
| person_female_first_cap.lst | Same as previous, but in upper case | person_first | female | 8380 |
| person_male_first.lst | Male first names from US census, merged with ANNIE's default list | person_first | male | 4788 |
| person_male_first_cap.lst | Same as previous, but in upper case | person_first | male | 4788 |
| surname_us_90_census.lst | Surnames from US census of 1990 | person_last | | 88799 |
| surname_us_90_census_cap.lst | Same as previous, but in upper case | person_last | | 88799 |
| Bush41whitehousestaff_full.lst | Full names of White House Staff of the Bush Administration | person_full | | 461 |
| Bush41whitehousestaff_full_cap.lst | Same as previous but in all caps | person_full | | 461 |
| US_presidents.lst | The full names of the 43 Presidents of the US plus some nicknames | person_full | | 61 |
| US_presidents_cap.lst | Same as US_Presidents, but in all caps, including F.D.R., J.F.K. and L.B.J. | person_full | | 64 |
| Senators_101_102_congress_names.lst | Full names of Senators in 101st & | person_full | | 107 |

| Wordlist | Description | Major Type | Minor Type | Count |
|---|---|---|---|---|
| | 102nd Congress | | | |
| Senators_101_102_congress_names_cap.lst | Same as previous, but in upper case | person_full | | 107 |
| House_101_102_congress_names.lst | Full names of Respresentives in 101st and 102nd Congress | person_full | | 435 |
| Heads_of_state_90.lst | Full names of heads of state and heads of government form the 1990 CIA World Factbook | person_full | | 474 |
| Heads_of_state_90_last_names.lst | Same as previous, but only last names | person_last | | 453 |
| Bush_Nominees_names.lst | Full names of Nominees to Federal Office during Bush 41 Administration. From Public Papers Appendix B | person_full | | 1354 |
| Bush_Appointments_names.lst | Appointment Announcements in Bush Public Papers, 1989-93. | person_full | | 123 |

**Figure 1. Additional Wordlists Used for Recognizing Persons Names**

## *2.2 Titles and Position Names*

It is important to be able to automatically recognize the job titles of Presidential Appointees other than White House Staff Members. Such a list was created from the announcements of Presidential Appointments published in the Federal Register and reprinted in the Bush Presidential Public Papers. There are additional appointee titles appearing in Appendix A of the Bush Public Papers.

It is important to be able to recognize the position names or job titles of Presidential Nominees that must be approved by the Senate. A list of these has been created from the list of Presidential Nominations in Appendix B of the Bush Public Papers.

The titles of Chiefs of State and Heads of Government must also be recognized. These were automatically extracted from the CIA World Factbook for 1990.

Twenty additional White House Staff titles were added to that wordlist. These were identified from announcements of appointments in the Bush Public Papers.

Fig. 2 summarizes the additions to word lists for titles and position names.

| Wordlist | Description | Major Type | Minor Type | Count |
|---|---|---|---|---|
| Bush_Appointments_titles.lst | Titles of Appointees for Bush Public Papers 1989-93 | jobtitle | | 77 |
| Bush_Nominees_titles.lst | Nominees to Federal Office requiring Senate approval, 1989-1992 | jobtitle | | 513 |
| Foreign_heads_of_state_govt_titles.lst | From CIA World Factbook 1990 | jobtitle | | 119 |
| Bush41White HouseStaffTitles.lst | White House staff titles | jobtitle | | 208 |

**Figure 2. Additional Wordlists Used for Recognizing Titles and Position Names**

## *2.3 Locations and Facilities*

In the previous experiment, there were 377 location names in the corpus. The semantic annotation method correctly identified 249 of these. However, 115 location names were missed, 13 of the identifications were only partially correct, and there were six false positives.

The list of city names included only 1969 names of cities worldwide. In this list, there are only 350 names of cities, towns and villages in the US. This accounts for the difficulties encountered in recognizing location names of US cities. This list has been replaced with a list of US city names containing 47,673 entries.

"The GEOnet Names Server GNS provides access to the National Geospatial-Intelligence Agency's (NGA) and the U.S. Board of Geographic Names (US BGN) database of foreign geographic feature names. The database is the official repository of foreign place-name decisions of the US BGN. The GNS contains approximately 4.0 million features with 5.5 million names."[2]

The United Nations *Demographic Yearbook* disseminates statistics on population size and composition, births, deaths, marriage and divorce on an annual basis. The Population of capital cities and cities of 100,000 and more inhabitants was downloaded in Excel spreadsheet format from the Demographic Yearbook 2004.[3] US City names were deleted from this spreadsheet leaving 3770 foreign city names.

"The Geographic Names Information System (GNIS) is the Federal standard for geographic nomenclature. The U.S. Geological Survey developed the GNIS for the U.S. Board on Geographic Names as the official repository of domestic geographic names data; the official vehicle for geographic names use by all departments of the Federal

[2] URL: http://earth-info.nga.mil/gns/html/index.html
[3] URL: http://unstats.un.org/UNSD/Demographic/products/dyb/dyb2004.htm

Government; and the source for applying geographic names to Federal electronic and printed products."

"The GNIS contains information about physical and cultural geographic features of all types in the United States, associated areas, and Antarctica, current and historical, but not including roads and highways. The database holds the Federally recognized name of each feature and defines the feature location by state, county, USGS topographic map, and geographic coordinates. Other attributes include names or spellings other than the official name, feature designations, feature classification, historical and descriptive information, and for some categories the geometric boundaries."[4]

Of particular use are:

- The physical and cultural geographic features for states, territories, and associated areas of the United States (updated Jan. 25, 2007).

- Topical Gazetteers of Populated Places (updated on Feb. 2, 2007) – Named features with human habitation—cities, towns, villages, etc. It contains 183,771 entries This file also includes county names.

County names also occur in the Bush e-records. The Topical Gazetteer of Populated Places contains named features with human habitation—cities, towns, villages, etc. It contains 183,771 entries. The file also includes county names. The county names were extracted from this file and duplicates eliminated.

In the previous experiment, "World Congress Center," "CNN Center" and "Omni Hotel" were recognized as entity type organization, when in the documents they are actually facilities. A wordlist facility.lst was created to handle the facility subtype of location. Rather than construct wordlists for all hotels, airports, and centers, one should create a wordlist, possibly called facility_suffixes.lst, that contains words such as Hotel, Airport and Center, and JAPE rules that recognize proper nouns proceeding these terms as constituting facility names.

Figure 3 summarizes additions to the wordlists for location and facility names.

| Wordlist | Description | Major Type | Minor Type | Count |
|---|---|---|---|---|
| Cities_US.lst | US Cities that have zip codes. | location | city | 47673 |
| Cities_US_cap.lst | Same as Cities_US.lst, but in all caps | location | city | 47673 |
| Foreign_cities.lst | Names of capitals of foreign countries and cities with population of 100,000 or more. | location | city | 3770 |
| Foreign_cities_cap.lst | Same as previous, but in upper case. | location | city | 3770 |
| US_counties.lst | Names of US counties | location | county | 1948 |
| US_counties_cap.lst | Same as previous, but in upper | location | county | 1948 |

[4] URL: http://geonames.usgs.gov/domestic/index.html

| Wordlist | Description | Major Type | Minor Type | Count |
|---|---|---|---|---|
| | case. | | | |
| facility.lst | Names of facilities such as airports, hotels and centers. | location | facility | 2 |

**Figure 3. Additional Wordlists for Locations and Facilities**

## *2.4 Organization Names*

There were 459 organization names in the corpus of the previous experiment. 279 organization names were correctly annotated, 46 were partially correct, 125 organization names were not recognized, and there were 67 false positives.

The transcripts of Presidential Press Conferences include the names of journalists and news media organizations. Among the organization names that were missed were the names of print media (newspapers and magazines) organizations, such as Dallas Morning News, Newsday and Newsweek; broadcast media, such as ABC News, CBS News, NBC News, CNN, and Fox News; media companies that own newspapers and/or broadcasting stations; and news agencies such as United Press International and the Associated Press. Wordlists of US newspapers,[5] US news magazines, broadcast media, media companies and news agencies were created.

US Departments and Agency names frequently occur in the Bush e-records. A list 450 US Government Departments and Agencies was found. The list included state department and agency names that were eliminated. Abbreviations for the departments were also entered in the wordlist.

Hospital names sometimes occur in the records. A list of US hospital names was created. University and college names often occur in biographies or resumes. A list of 1951 US colleges and universities was created.

Terms that are names of organizations, such as universities, colleges and hospitals, may also refer to facilities and thus need to be disambiguated. This issue is discussed in sections 2.3 and 3.3.

Figure 4 summarizes additions to organization name wordlists.

| Wordlist | Description | Major Type | Minor Type | Count |
|---|---|---|---|---|
| newspapers_US.lst | | organization | news_media | 2694 |
| newspapers_US_cap.lst | Same as newspapers_US.lst but capitalized | organization | news_media | 2694 |
| | | organization | news_media | 11 |
| | | organization | news_media | 3 |
| | | organization | news_media | 5 |
| | | organization | mews_media | 7 |

[5] URL: www.50states.com/news/

| Wordlist | Description | Major Type | Minor Type | Count |
|---|---|---|---|---|
| US_government.lst | Names of US government departments and agencies. | organization | government | 450 |
| US_government_cap.lst | Same as US_government.lst but capitalized | organization | government | 450 |
| US_hospital.lst | Names of US hospitals | organization | hospital | 4168 |
| US_hospital_cap.lst | Same as US_hospital.lst, but all caps | organization | hospital | 4168 |
| US_college_university.lst | Names of US colleges and universities. | organization | educational | 1951 |
| US_college_university_cap.lst | Same as US_college_university.lst, but all caps | organization | educational | 1951 |

**Figure 4. Additions to Organization Name Wordlists**

## *2.5 Money*

In the previous experiment, American monetary terms such as penny, nickel, dime, quarter, and bucks were not recognized. They have been added to the currency_unit.lst.

## *2.6 Relative Temporal Expressions*

There are 281 date/time expressions in the second experimental corpus. Of these, 268 were correctly annotated, one was partially correct, and 12 were missed. There were 17 false positives. Among those missed were:

The year of our Lord nineteen hundred and ninety
Second day of October
1976
4-year Limitation
2/28
six months ago
WITHIN 9 WORKING DAYS
Midnight EDT on April 25
8 years
last few days
4 days of this week
weekend
This weekend
Two weeks ago
present

Relative temporal expressions, such as "Day after tomorrow," need to be recognized in order to understand sequence of actions or events mentioned in the text. In order to

recognize relative temporal expressions, the vocabulary needed to be expanded beyond typical date units such as day, week, and hour into adverbs and adjectives, such as last, couple, early, and late, that modify these time words. To this end, a basic time modifiers list (time_mods2.lst) was created that includes words such as the following:

few
last
this
present
current
past
couple
now
early
late

Time expressions also use *frequency* words such as often, only, once, and twice. Frequency terms indicate the number of occurrences an event in a time period. Consider the example: “We met just once in the past couple of days”. In earlier implementations, the word ‘days’ would just be annotated as a date. With frequency in consideration, ‘just once in the past couple of days’ is the complete time expression. A wordlist of general time-frequency words was created with the major type being time modifier and minor type being frequency.

The following are general frequency terms (time_frequency.lst).

even
further
just
only
more
less
most
nearly
once
twice
sometimes

The following are adverbs indicating the frequency of an event (advFreq.lst).

usually
periodically
repeatedly
continuously
endlessly
always
hardly

barely
seldom

The following are adjective prefixes to the frequency of an event (adjFreqPrefix.lst).

each
every

The following adverbs indicate the degree of the frequency of an event (advFreqDegree.lst).

even
further
just
only
more

The following are examples of adjectives that indicate the frequency of an event (adjFreq.lst).

endless
periodic
recurrent
steady
rare
bare
scarce

Figure 5 summarizes additions to the wordlists to facilitate recognition of relative temporal expressions.

| Wordlist | Description | Major Type | Minor Type | Count |
|---|---|---|---|---|
| AdjFreq.lst | Frequency Adjectives | time | frequency | 13 |
| AdjFreqPrefix.lst | Frequency adjectives | time | frequency | 2 |
| AdvFreq.lst | Frequency adverbs | time | frequency | 17 |
| AdvFreqDegree.lst | Degree of Frequency adverbs | time | frequency | 10 |
| time_key.lst | General time modifiers | time_modifier | | 12 |
| time_mods2.lst | General time modifiers | time_modifier | | 13 |
| Time_suffix.lst | Words found at the end of time expressions, e.g., ago | time_suffix | | 1 |
| Timex_pre.lst | Modifier that precede time words | time_modifier | | 26 |

**Figure 5. Additions to Wordlists for Relative Temporal Expressions**

## *2.7 Congressional Bills and Statutes*

In the previous experiment, the abbreviation "H.R." for House Resolution was incorrectly recognized as the initials of a person's name and the resolution number was mis-annotated as a date. Lists of Bills of the 101st Congress [LOC 2007] were created as well as lists of typical bill and act prefixes. However, after the rules were written based on the bill types and bill prefixes, the lists of resolutions and statutes are not necessary in determining bills and acts. In this way, recognition is not dependent on exhaustive lists that would need updating as more statutes are added.

| Wordlist | Description | Major Type | Minor Type | Count |
|---|---|---|---|---|
| bills_101st_congress.lst | Names of Bills of the 101st Congress | bill | name | 987 |
| bill_type.lst | Names of types of congressional bills, e.g., House Resolution | bill | type | 8 |
| bill_type_abbr.lst | Abbreviations for types of congressional bills, e.g., H.R., S. | bill | type | 27 |
| abbr_congressional_bills_101st.lst | Abbreviations of congressional bills of 101st congress with version, e.g. H.R.2362.IH | bill | abbr | 1000 |
| abbr_bills_101st_sans_version.lst | Abbreviations of congressional bills of 101st congress without version, e.g., H.R.2362 | bill | abbr | 1000 |

# 3. Improvements to JAPE Rules

## *3.1 Rules Added to Correct Annotation Errors*

### 3.1.1 Job Titles as Person's Names

People are specified by name (“George Bush”), position or job title (“the President”), family relation (“dad”), or pronoun (“he”). When the article "the" precedes a job title such as "the President", the job title will be considered to be the name of a person. Other examples are "the Queen", "the General," "the Major," the Chief of Staff," and "the Secretary of State."

```
Rule: theJobTitle
Priority: 20
// the President
// the Queen
(
 {Token.string == "the"}
```

```
 {Lookup.majorType == jobtitle}
)
:person
-->
   :person.TempPerson = {kind = "personName", rule = "theJobTitle"}
```

### 3.1.2 Job Titles that Include Office Names

Organization names within job titles are annotated as organization names. However, the complete job title includes the organization name. For instance, the job title "Assistant to the President for Legislative Affairs" includes the office "Legislative Affairs." Rules for recognizing the job titles that include office names have been created.

### 3.1.3 Names of Law Firms

The names of law firms are typically a sequence of partner's (person's) last names with an ampersand before the name of the last partner. A JAPE rule was created for recognizing this pattern and annotating it with major type *organization* and minor type *law_firm*.

```
Rule: LawFirm
Priority: 205
// lawfirm: list of names
(
  (({FIRSTNAME} | {Token.orth == upperInitial} | {Token.category == NNP})
  {Token.string == ","}
  )+
    ({FIRSTNAME} | {Token.orth == upperInitial} | {Token.category == NNP})
    {Token.string == "&"}
      ({FIRSTNAME} | {Token.orth == upperInitial} | {Token.category == NNP})
)
:orgName -->
   :orgName.TempOrganization = {kind = "lawfirm", rule = "LawFirm"}
```

### 3.1.4 Social Security Numbers

JAPE rules were added to recognize social security numbers in text. First, a macro was created to recognize words that precede social security numbers such as: Social Security Number, SSN#, SSN#:, SSN: combinations. Then, two rules were created to search for social security numbers in the format of: 444-44-4444 or 9 digit number sequences. Rules are shown below.

```
Macro: SSN_PRE
(
{Token.string == "SSN"} |
```

```
{Token.string == "ssn"} |
{Token.string == "Social"}
{Token.string == "Security"}
{Token.string == "Number"} |
{Token.string == "SSN"}
{Token.string == "#"} |
{Token.string == "ssn"}
{Token.string == "#"} |
{Token.string == "SSN"}
{Token.string == ":"} |
{Token.string == "ssn"}
{Token.string == ":"} |
{Token.string == "SSN"}
{Token.string == "#"}
{Token.string == ":"}

Rule: SSN_Number
// 555-55-5555
// SSN: 555-55-5555
// SSN: 555555555
(
(SSN_PRE)?
(THREE_DIGIT)
{Token.string == "-"}
(TWO_DIGIT)
{Token.string == "-"}
(FOUR_DIGIT)
)
:date -->
 :date.SS_Num = {kind = "fullNumber", rule = "SSN_Number"}
Rule: SSN_Number2
// SSN: 555555555
(
(SSN_PRE)
(NINE_DIGIT)
)
:date -->
 :date.SS_Num = {kind = "fullNumber", rule = "SSN_Number2"}
```

### 3.1.5 Rules for Recognizing Money Terms

Some JAPE rules were added to enhance recognition of the semantic category money. Previously, there was a moneysymbolunit rule that required a symbol like $ to prefix ‘30 million bucks’ or ‘8 billion dollars.’ However, monetary units often occur without the dollar sign. A moneyunit rule was added.

```
Rule:   MoneyUnit
// 30 million
// 30 million bucks
(
 (AMOUNT_NUMBER)?
 {Lookup.majorType == currency_unit, Lookup.minorType = post_amount}
)
:number
 -->
  :number.Money = {kind = "number", rule = "MoneyUnit"}
```

The ? option was added to amount_number in order to recognize ‘million bucks.’

### 3.1.6 Rules for Postal Addresses

The US Postal Service defines *address* as "The location to which the USPS is to deliver or return a mail piece. It consists of certain elements such as recipient name, street name and house number, and city, state, and ZIP Code as required by the mail class [USPS 1997]." Hence, the recipient name was included in the address annotation.

## *3.2 Distinguishing Names of Countries from the Names of Their Governments*

Previously, terms such as the *United States, America* and *England* were annotated as a location of subtype country. However, these terms are ambiguous. Depending on context, they can refer to the geographic area, the government (a political entity) or the populace (citizens) A similar situation holds for the names of cities.

The *UNITED STATES OF AMERICA* in *PRESIDENT OF THE UNITED STATES OF AMERICA* is a government, not a location, because the President is not the President of a location, but of a government. *PRESIDENT OF THE UNITED STATES OF AMERICA* as a whole is also a job title or position.

The term *China* in the sentence "What signal do you think it may send the world, Sir, that you're making your first visit to China?" refers to a location of type country.

The phrase *Peoples Republic of China* in the sentence "We have a strong bilateral relationship with the People's Republic of China." is the name of an organization of subtype government.

The phrase *Americans with handicaps* in the sentence

> "In order to fulfill President Bush's Campaign promise of bringing Americans with handicaps into the mainstream of American life, the Bush Administration supports the objectives of the A.D.A."

refers to a person of subtype group.

The Automatic Content Extraction (ACE) Conferences addressed this issue by creating an artificial entity termed Geographical-Political Entity (GPE) that subsumed these three concepts. The subtypes are termed roles [LDC 2006].

Our approach to distinguishing these three concepts is to use wordlists to initially annotate country and city names as locations of subtype country or city. Then to use JAPE rules to disambiguate the term by referring to the context of the term.

When the term refers to a geographic area, it is annotated as a location of subtype country. When it refers to a government, it is annotated as an organization of subtype government. When it refers to citizens, it is annotated as person of subtype group.

## *3.3 Distinguishing Organization Names from Facilities with the Same Name*

Phrases, such as *The White House*, need to be disambiguated as to whether they refer to a facility (the building) or to a government such as the President and the White House Staff. Such a phrase is annotated as type organization of subtype government when the textual reference is to the organization, and as type location, subtype facility when the reference is to the physical building.

> Members of the Press met in the Pressroom of the White House. The White House Press Secretary responded to their questions.

The first reference to *White House* is to the physical building. The second reference is to an organization of subtype government.

In document 129 of the experimental corpus, the following sentence occurs.

> I’ve been reading your press and receiving a lot of complaints from the Hill.

In this case, "the Hill" is "Capitol Hill", the common nickname for the United States Congress. Hence, "the Hill" is an organization or a group of persons. In the same document, "The Kitchen" is a theatre company and a theater in Manhattan that has featured explicit acts by feminist and homosexual performers. Whether "The Kitchen" refers to an organization or a facility depends on the context.

If text is "They met at Georgia Tech.", then Georgia Tech is a location of minor type facility, i.e., the Georgia Tech Campus. This can be recognized by the preposition "at" preceding Georgia Tech. In a resume or biography, the reference to an educational institution is to the organization that granted the degree, not to a facility. If a city name is mentioned with the name of the college, the city name is the location (ogf subtype facility) where the institution is located. The following JAPE rule is used to recognize names of educational or hospital facilities.

```
Rule: OrgFacility_UnivHosp
Priority:205
(
  {Token.string == "at"}
  (
  {Lookup.majorType == organization, Lookup.minorType == educational} |
  {Lookup.majorType == organization, Lookup.minorType == hospital}
  )
)
-->
:orgName.TempOrganization = {kind = "facility", rule=OrgFacility_UnivHosp}
```

## *3.4 Rules for Relative Temporal Expressions*

JAPE rules were written to recognize temporal expressions. Frequency words were not annotated on their own. They were only recognized in conjunction with their proximity to the more traditional time-unit and time-unit modifier phrases.

Below are some examples of time expressions that are now recognized using JAPE rules. The heading of each table indicates the pattern that was used to recognize the time expression. For instance, in the first table, a time modifier followed by a determiner (DT), followed by a time-unit is a time expression. .

| (Time modifiers)? | <DT> | time-unit |
|---|---|---|
| Early | this | year |
| | this | weekend |

| (Time modifiers)+ | (<IN DT>)? | time-unit |
|---|---|---|
| Early | in the | day |
| Last few | | years |

| <Date> | (<TO DT>)? | <Date> |
|---|---|---|
| Year of our Lord | | 1999 |
| 1997 | to the | present |

| (number \| number words) | time-unit | (<DT> \| <IN DT>) | time-unit |
|---|---|---|---|
| 4 | days | this | week |
| Second | day | of the | week |

The following rule is used to recognize the temporal expressions in the previous table.

```
Rule: TimePhrase3
 // 4 days this week or second day of the week
 // (number | number words) time-unit (<DT> | <IN DT> ) time-unit
 (
 (NUM_OR_ORDINAL)
 {Lookup.majorType == time}
 ({Token.category == DT} | {Token.category == IN} {Token.category == DT} )
 ({Lookup.majorType == time_modifier})?
 {Lookup.majorType == time}
 ):time
 -->
 :time.TempTime = {kind = "timePhrase", rule = "TimePhrase3"}
```

## *3.5 Rules for Congressional Bills and Statutes*

Presidential records created in the Office of Legislative Affairs often refer to Congressional Bills (e.g., House Resolutions) and Statutes or Acts (e.g., Americans with Disabilities Act). These bills and statutes need to be automatically identified in text and annotated. The following definitions are from the Congressional Bills Glossary [GPO 2006].

"A bill is a legislative proposal before Congress. Bills from each house are assigned a number in the order in which they are introduced, starting at the beginning of each Congress (first and second sessions). Public bills pertain to matters that affect the general public or classes of citizens, while private bills pertain to individual matters that affect individuals and organizations, such as claims against the Government."

"A joint resolution is a legislative proposal that requires the approval of both houses and the signature of the President, just as a bill does. Resolutions from each house are assigned a number in the order in which they are introduced, starting at the beginning of each Congress (first and second sessions). There is no real difference between a bill and a joint resolution. Joint resolutions generally are used for limited matters, such as a single appropriation for a specific purpose. They are also used to propose amendments to the Constitution. A joint resolution has the force of law, if approved. Joint resolutions become a part of the Constitution when three-quarters of the states have ratified them; they do not require the President's signature."

"A concurrent resolution is a legislative proposal that requires the approval of both houses but does not require the signature of the President and does not have the force of law. Concurrent resolutions generally are used to make or amend rules that apply to both houses. They are also used to express the sentiments of both of the houses. For example, a concurrent resolution is used to set the time of Congress' adjournment. It may also be

used by Congress to convey congratulations to another country on the anniversary of its independence."

"A simple resolution is a legislative proposal that addresses matters entirely within the prerogative of one house or the other. It requires neither the approval of the other house nor the signature of the President, and it does not have the force of law. Most simple resolutions concern the rules of one house. They are also used to express the sentiments of a single house. For example, a simple resolution may offer condolences to the family of a deceased member of Congress, or it may give "advice" on foreign policy or other executive business."

"A report is a document that presents a committee's explanation of its action regarding legislation that has been referred to it. Each House and Senate report is assigned a number that includes the number of the Congress during which it is published (e.g., "H.Rpt. 105-830" refers to a report created in the House during the 105th Congress). Conference reports are numbered and designated in the same way as regular House and Senate reports. Most reports favor a bill's passage, although a bill can be reported without recommendation. When a committee report is not unanimous, the dissenting committee members may file a statement of their views (minority views) in a minority report. A reported version of a bill references the applicable report number."

An *act* is "Legislation (a bill or joint resolution) which has passed both chambers of Congress in identical form, been signed into law by the President, or passed over his veto, thus becoming law. Technically, this term also refers to a bill that has been passed by one house and engrossed (prepared as an official copy)." [6]

A *public law* is "A public bill or joint resolution that has passed both chambers and been enacted into law." [7]

The following JAPE rule takes bill and statute prefixes like HR, S and S.Res and looks for following multiple digits. This rule would recognize HR 2456 or S. Res 2345.

```
Rule: BillPrefix
Priority: 250
//takes bill and statute prefixes like HR and S and S.Res and looks for following digits
(
{Lookup.majorType == bill, Lookup.minorType == prefix}
({Token.kind == number})+
)
:bill --> :bill.billName = {kind = "billName", rule = BillPrefix}
```

The following JAPE rule annotates the names of acts by looking for multiple uppercase words preceding the word Act.

[6] United States Senate Glossary URL:
http://www.senate.gov/pagelayout/reference/b_three_sections_with_teasers/glossary.htm
[7] *ibid.*

```
Rule: UpperInitialAct
Priority:250
//Job Training Partnership Act
(
 (UPPER)*
 (
  (UPPER)|
  {Lookup.majorType == organization}
 )
 (UPPER)*
{Token.string == "Act"}
 (
          ({Token.string == "Amendments"})?
          {Token.category == IN}
          {Token.kind == number}
    |
          {Token.string == ","}
          {Token.kind == number}
 )?
)
:bill --> :bill.Act = {kind = "Act", rule = UpperInitialAct}
```

The following JAPE rule recognizes *lists* of uppercase words preceding the word Act and annotates them and the word *Act* as an Act.

```
Rule: UpperInitialActList
Priority:255
//Financial Institutions Reform, Recovery, and Enforcement Act of 1989
(
 ((UPPER)+
 {Token.string == ","})?
 (UPPER)+
 {Token.string == ","}
 (UPPER)+
 ({Token.string == ","})?
 {Token.category == CC}
 (
  (UPPER)|
  {Lookup.majorType == organization}
 )
 (UPPER)*
{Token.string == "Act"}
 (
          ({Token.string == "Amendments"})?
          {Token.category == IN}
```

```
            {Token.kind == number}
        |
            {Token.string == ","}
            {Token.kind == number}
     )?
    ) :bill --> :bill.Act = {kind = "Act", rule = UpperInitialActList}
```

# 4. Tests to Verify Improvements in Semantic Annotation

## *4.1 Verifying improvements in the Performance of the Semantic Annotation of Corpus 2*

The results of the experiment that applied the PERPOS information extractor to the second corpus is shown below.

| Annotation Type | Correct | Partially Correct | Missing | Spurious | Precision | Recall | F-Measure |
|---|---|---|---|---|---|---|---|
| Person | 333 | 66 | 64 | 145 | 0.6728 | 0.7905 | 0.7269 |
| Location | 249 | 13 | 115 | 6 | 0.9534 | 0.6777 | 0.7922 |
| Organization | 279 | 46 | 124 | 67 | 0.7704 | 0.6726 | 0.7182 |
| Date | 268 | 1 | 12 | 17 | 0.9388 | 0.9555 | 0.9471 |
| Money | 19 | 1 | 2 | 0 | 0.975 | 0.8864 | 0.9286 |
| Percent | 24 | 3 | 0 | 1 | 0.9107 | 0.9444 | 0.9273 |

Overall average precision: 0.8582. Overall average recall: 0.8066 F-measure: 0.8316

It was discovered that there were some inconsistencies in how the JAPE rules were annotating text, and how the human who constructed the key files was annotating text. Furthermore, there are additional semantic categories that are being annotated, e.g., telephone numbers, social security numbers, Congressional Bills and Statutes. Hence, the Key files were re-annotated.

The additional and modified wordlists and the modifications to the JAPE rules have been tested on the second experimental corpus. The results are shown below.

| Annotation Type | Correct | Partially Correct | Missing | Spurious | Precision | Recall | F-Measure |
|---|---|---|---|---|---|---|---|
| Person | 365 | 34 | 44 | 52 | 0.847 | 0.8623 | 0.8546 |
| Location | 289 | 13 | 55 | 14 | 0.9351 | 0.8277 | 0.8782 |
| Organization | 385 | 44 | 26 | 30 | 0.8867 | 0.8945 | 0.8906 |
| Date | 260 | 5 | 12 | 17 | 0.9309 | 0.9477 | 0.9392 |
| Money | 18 | 2 | 1 | 0 | 0.95 | 0.9048 | 0.9268 |
| Percent | 27 | 1 | 0 | 0 | 0.9821 | 0.9821 | 0.9821 |

Overall average precision: 0.9054049426838151. Overall average recall: 0.886117308270068

In this particular test, average precision rose from 0.858 to 0.905 and average precision from .0.806 to 0.886.

## *4.2 Verifying the Recognition of the Names of Heads of State/Government, Titles, and Countries*

The names of Chiefs of State and Heads of Government extracted from the CIA World Factbook for 1990 [Harris and Underwood 2004, Appendix N] was used as a test corpus to evaluate the performance of the modifications to JAPE Rules and wordlists for recognizing the names of Chiefs of state/Heads of Government, their titles and the names of countries. Figures 6 and 7 show excerpts from the results of the test. All names and titles in the test corpus were correctly recognized and annotated. Text highlighted in blue are job titles, in red are person's names, and in purple are locations of subtype country.

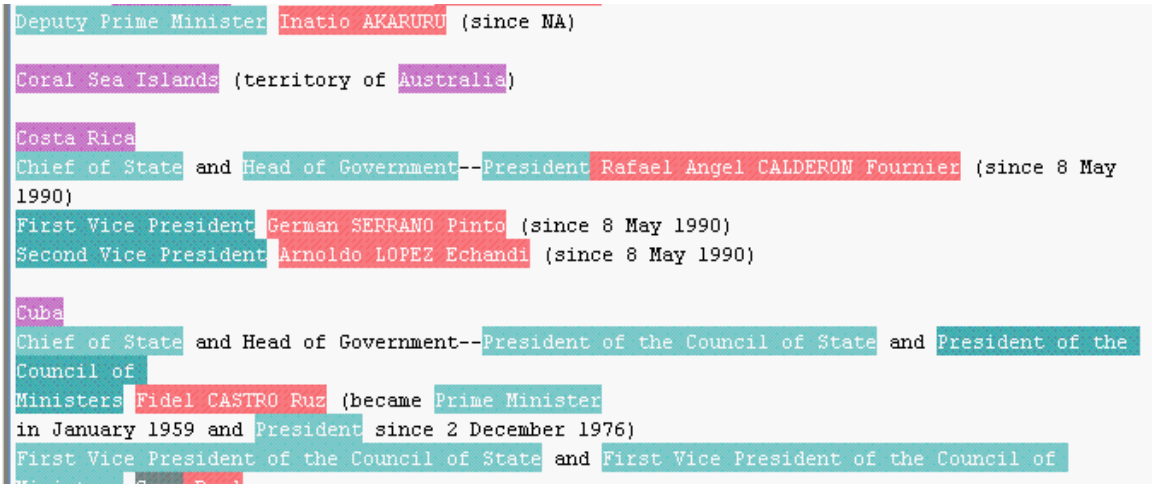


**Figure 6. Test of Annotation of Names of Chiefs of State/Heads of Government and their Titles.**

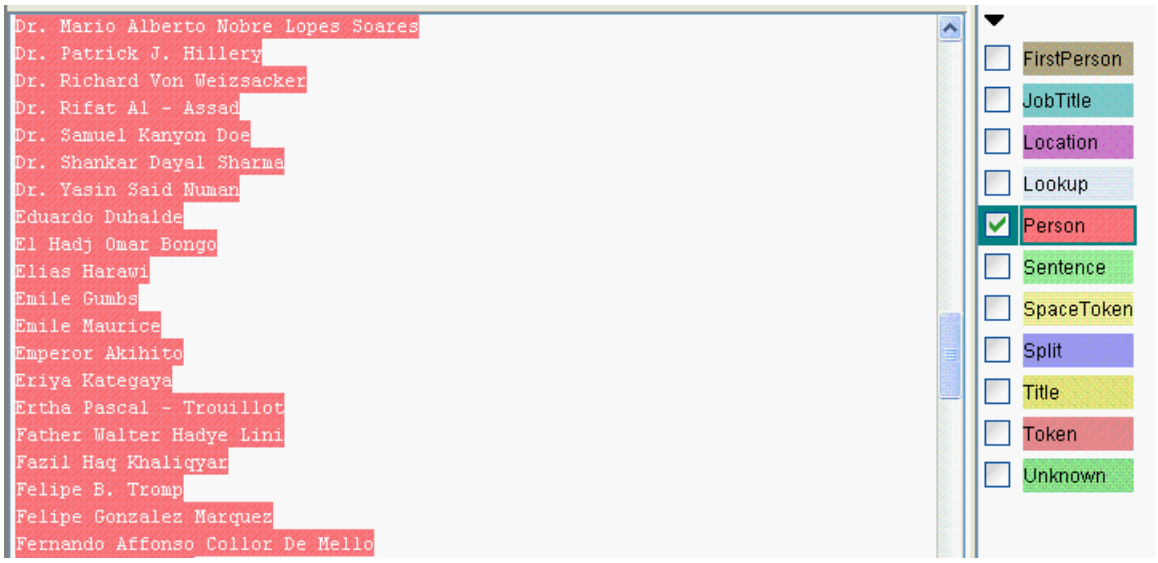


**Figure 7. Test of Annotation of Names of Chiefs of State/Heads of Government as Persons.**

## *4.3 Verifying the Recognition of the Names and Titles of Presidential Nominees and Appointees*

Appendix B of the Bush Presidential papers includes the names and titles of Presidential nominees to Federal Office that require the consent of the Senate. Appendix A includes the names and titles of many Presidential appointees that do not require approval of the Senate.

A test was conducted to ensure that additions to the wordlists and modifications to the JAPE rules for person's names and position titles were effective in recognizing Nominee and Appointee names and titles. Figure 7 shows an excerpt of the results of the test. All names and titles in the test corpus were correctly recognized and annotated.

Ambassador Extraordinary and Plenipotentiary of the United States of America to Papua New Guinea, and to serve concurrently and without additional compensation as Ambassador Extraordinary and Plenipotentiary of the United States of America to Solomon Islands, and Ambassador Extraordinary and Plenipotentiary of the United States of America to the Republic of Vanuatu.

J. Steven Rhodes,

of California, to be Ambassador Extraordinary and Plenipotentiary of the United States of America to Zimbabwe.

John J. Maresca,

of Connecticut, a career member of the Senior Foreign Service, Class of Minister-Counselor, for the rank of Ambassador during his tenure of service as Head of the United States Delegation to the Conference on Confidence and Security Building Measures (CSBM).

Carol Mayer Marshall,

of California, to be Superintendent of the Mint of the United States at San Francisco, California (new position).

Jerome G. Cooper,

of Alabama, to be an Assistant Secretary of the Air Force, vice Karen R. Keesling, resigned.

Rhesa H. Barksdale,

of Mississippi, to be United States Circuit Judge for the Fifth Circuit, vice Alvin B. Rubin, retired.

Jacques L. Wiener, Jr.,

of Louisiana, to be United States Circuit Judge for the Fifth Circuit, vice Robert M. Hill, deceased.

Ronald L. Buckwalter,

of Pennsylvania, to be United States District Judge for the Eastern District of Pennsylvania, vice Charles R. Weiner, retired.

**Figure 8. Excerpt from the Test of the Recognition of the Names and titles of Presidential Nominees.**

## *4.4 Verifying the Recognition of Bills and Acts*

Appendix D of the Bush Public Papers lists Bills of the 101st and 102nd Congress approved by the President, or passed by Congress over his veto, thus becoming an Act.. These were used as a test corpus for the changes made to recognize bills and statutes. Figure 8 shows an excerpt from the test of the recognition of Congressional bills and statutes. All bills and statutes in the test corpus were correctly recognized and annotated.

```
District of Columbia Civil Contempt Imprisonment Limitation Act of 1989

Approved September 26

H.J. Res. 133 / Public Law 101 - 98

Designating the week beginning September 17, 1989, as ``Emergency Medical Ser

S. 1075 / Public Law 101 - 99

To authorize appropriations for the American Folklife Center for fiscal years
1992

Approved September 29

H.J. Res. 407 / Public Law 101 - 100

Making continuing appropriations for the fiscal year 1990, and for other purp

H.R. 2696 / Public Law 101 - 101

Energy and Water Development Appropriations Act, 1990

H.J. Res. 204 / Public Law 101 - 102

To designate October 1989, as ``National Quality Month''

Approved September 30

H.R. 3282 / Public Law 101 - 103

Performance Management and Recognition System Reauthorization Act of 1989

Approved October 2

S.J. Res. 146 / Public Law 101 - 104
```

**Figure 9. Excerpt from the Test of the Recognition of Congressional Bills and Statutes.**

# 5. Scalability of the Semantic Annotation Methods

Tests were conducted to estimate the throughput of the Information Extraction System. The modified ANNIE Information Extraction system with 137 gazetteers containing 60,302 terms and 211 JAPE rules and 53 macros was used to markup person, organization and location names, dates, money and percents in text versions of the Bush Public Papers. There are 5,173 text documents in the Bush Public Papers with a total size of 28,578,968 bytes. The largest file is 99,410 bytes and the smallest file 373 bytes. The average size is 5524 bytes.

The Default ANNIE configuration uses a finite state transducer to lookup terms in the gazetteer and to apply the JAPE rules to annotate the named entities. Running under Windows 2000 on a Pentium 4 with a 3.06 GHZ processor, ANNIE marked up the documents in 111 minutes, 58 seconds. This amounts to processing 4254 bytes per second. Given an average size of 5524 bytes, ANNIE is processing a document of that size in 1.3 seconds.

One of the strengths of the ANNIE architecture is the capability to substitute processing and language resources. A test was conducted by substituting the Ontotext Hash Gazetteer for the finite state transducer that looks up terms in the wordlists. The Ontotext Hash gazetteer purportedly uses 4 times less memory while working three times faster. The 5,173 documents were processed in 106 minutes 14 seconds. This amounts to 4519 bytes per second. Given an average size of 5524 bytes, this improved configuration of ANNIE is processing a document of that size in 1.22 seconds.

It is anticipated that this semantic annotation technology could be used to process millions of records. On an Intel 3 GHZ processor, it takes better than an hour to annotate 5000 documents. To determine whether the technology scales to such volumes, the JAVA code for the semantic annotator should be ported to a supercomputer supporting some dialect of JAVA, such as Titantium,[8] and experiments conducted to determine the throughput.

# 6. Experiments in Semantic Annotation of Bush Administration Personal Computer Records

During April, tools for extracting files from containers, for converting the files to text formats, and for annotating their contents were installed on the PERPOS system in the Virtual Laboratory at Archives II. They will be used in experiments to evaluate the performance of the semantic annotation of actual, born digital, personal computer records from the Bush Administration. This is also a first step toward interfacing these tools to PERPOS to support document type learning, document type recognition, metadata extraction and automatic description of items, folder and record series.

[8] http://titanium.cs.berkeley.edu/

A corpus of 50-100 records will be selected from the record series accessioned into PERPOS. They will most likely be records from the Office of Legislative Affairs and the Office of Presidential Personnel. The Office of Legislative Affairs provides advice and support regarding the President's legislative agenda and legislation is general, and liaison between the White House staff and members of Congress. The Office of Presidential Personnel is responsible for reviewing all applicants for Presidential Boards and Commissions, senior staff positions in the White House and Cabinet Offices, and ambassadorships.

There will be two experiments. The first experiment will evaluate the performance of the semantic annotator with regard to the named entities addressed in the previous two experiments, namely, annotation of person, location, and organization names, dates, money and percents. The second experiment will address the annotation of additional semantic categories, namely, job titles, social security numbers, postal addresses, facilities, congressional bills and statutes, relative temporal expressions and organizations of type government and persons of subtype group.

# 7. Summary of Results

The results of a previous information extraction experiment were analyzed to determine reasons for partially correct annotatons, missing annotations and false positives. The solutions to the problems were in the provision of additional wordlists for primitive semantic categories and additional or modified JAPE rules.

Additional wordlists and JAPE rules were created for recognizing and annotating new semantic categories, namely, relative temporal expressions, and congressional bills and statutes. Additional JAPE rules were constructed for disambiguating terms that might designate either locations or facilities, and disambiguating terms that might designate countries as geographic areas, governments of countries and citizens of countries.

Previously, experiments were conducted on copies of paper Presidential records that were scanned and OCRed. The current experiments will be conducted using e-records from the Bush Presidential personal computer records. Two experiments will be conducted. The first experiment will evaluate the performance of the semantic annotator with regard to the named entities addressed in the previous two experiments, namely, names of persons, locations and organizations, and dates, money and percents. The second experiment will address the annotation of new semantic categories, namely, job titles, social security numbers, postal addresses, facilities, congressional bills and statutes, relative temporal expressions and organizations of subtype government and persons of subtype group.

# References

[CIA 1990] CIA World Fact Book Electronic Version
http://manybooks.net/titles/usciaetext93world192.html

[GPO 2006] GPO Access. Congressional Bills: Glossary.
www.gpoaccess.gov/bills/glossary.html

[LOC 1990] Congressional Record 101st Congress (1989-1990)
http://thomas.loc.gov/home/r101query.html

[LOC 1992] Congressional Record 102nd Congress (1991-1992)
http://thomas.loc.gov/home/r102query.html

[Harris and Underwood 2004] B. Harris and W. Underwood. Factual Knowledge Needed for Information Extraction and FOIA Review. PERPOS Technical Report ITTL/CSITD 04-7 Georgia Tech Research Institute, Atlanta, Georgia, December 2004

[Isbell et al 2006] 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.

[LDC 2006] Linguistic Data Consortium. ACE (Automatic Content Extraction) English Annotation Guidelines for Entities. Version 5.6.6. August 1, 2006.
http://www.ldc.upenn.edu/Projects/ACE/

[LOC 2007] The Library of Congress. THOMAS.
URL: http://Thomas.loc.gov/home/bills_res.html

[USPS 1997] US Postal Service. Glossary of Postal Terms, Publication 32, May 1997
http://www.usps.com/cpim/ftp/pubs/pub32/pub32tc.htm