Semantic Annotation of Presidential E-Records

William Underwood
Sheila Isbell

Technical Report ITTL/CSITD 08-01

May 2008

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

The capability to extract metadata from 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 solutions to the problems were in the provision of additional wordlists for semantic categories and in additional or modified rules for annotating these semantic categories. Additional wordlists and rules have been created for recognizing and annotating additional semantic categories such as relative temporal expressions, and congressional bills and statutes.

An experiment was conducted to 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, and dates, money and percents. Previously, experiments were conducted on copies of paper Presidential records that were scanned and OCRed. The current experiment was conducted using records from the Bush Presidential personal computer records. The results are an overall average precision of 0.9178, overall average recall of 0.9282, and overall F-measure of 0.9108.

Table of Contents

Table of Figures

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 evaluated [Isbell et al 2006]. This technology is based on annotating text based on terms that appear in 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. Rules then apply to these annotations to produce more complex annotations such as person's full names, organization names, dates, and postal addresses.

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 as social security numbers and congressional bills and acts. The purpose of this report is to describe the additions to the wordlists and changes to the rules, and to discuss an experiment performed to evaluate the performance of the enhancements of the semantic annotation system.

1.3 Scope

In section 2, our approach to automatically annotating semantic categories in textual records is reviewed. In section 3, extensions to the wordlists are discussed. In section 4, improvements to the Java Annotation Pattern Engine (JAPE) rules are discussed. In

section 5, tests to verify improvements are described. In section 6, experiments in annotating Presidential e-records are discussed. In section 7, results are summarized.

2. Annotation of Semantic Categories

Our approach to annotating named entities in records is based on the General Architecture for Text Engineering (GATE) architecture [Cunningham et al 2007]. Furthermore, it uses processing and lexical resources, provided with the GATE distribution that is referred to as the ANNIE (A Nearly New Information Extraction) System. These resources are shown below:

Document Reader
Tokenizer
Wordlist Lookup + Wordlists
Sentence Splitter
Hepple POS Tagger + Lexicon
Named Entity Transducer + JAPE rules

In our research, we refer to named entities (person names, organization names, location names, dates) as semantic categories. They are to be distinguished from syntactic categories such as nouns, verbs and adjectives.

A previous experiment has shown that the wordlists and JAPE rules provided with the GATE distribution are adequate to demonstrate the method of semantic annotation, but inadequate for practical application [Underwood 2004]. This report describes improvements made to these two resources and an experiment to evaluate the performance with these enhanced resources.

3. 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.

3.1 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 errors were in large part due to person’s first and last names that appeared in the corpus, but were not included in wordlists.

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

Presidential records discussing appointments to Federal office contain the names of candidates and those selected for nomination or appointment. The Bush Public Papers contain the names and position titles of the persons nominated for Federal positions that require approval of the Senate and the names and job titles of Presidential appointees that do not require Senate approval. These names and titles were extracted from the Bush Public Papers for use in the wordlists.

Presidential records that are memoranda or correspondence frequently contain the names and titles of White House Staff, not all of whom are Presidential appointees. The names and titles of White House Staff were extracted from a copy of the White House telephone book.

The US Census Bureau 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. Last names were extracted from the lists described in the preceding paragraphs (Senators and Representatives, Nominees and Appointees, and White House Staff) and merged with the surname list from the US Census.[1] The resulting list contains 90,607 surnames.

Some terms such as Angel that are surnames may also be city names, or capitalized nouns that are not proper nouns. For instance,

> Robert Angel, Associate Professor in the Department of Government and International Studies at the University of South Carolina.
>
> Angel, Alabama (Angel, Ohio; Angel, Michigan)
>
> Angel – (1) a typically benevolent celestial being that acts as an intermediary between heaven and Earth; (2) a person having qualities generally attributed to an angel; (3) the US Secret Service code name for Air Force One.

Consequently, the Surname wordlist is separated into two lists, one with unambiguous surnames (83,805) names and the other with ambiguous surnames (6,802). The terms in both lists are unambiguous if they follow a person's first name or are followed by a state name or abbreviation.

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.[2] These names occurred in 90% of the US population. These two lists were merged with the first names of female congressmen, white house staff and presidential nominees and appointees producing a list of 8440 female first names.

[1] Except for the term "Unavailable" which appeared in the US Census surname list.
[2] http://www.census.gov/genealogy/names/names_files.html

Female first names may also be ambiguous. Angel is a female as well as male first name. Consequently, the person_female_first wordlist was compared with a dictionary of English terms and US city names to produce two lists, one with unambiguous female first names (8052 names) and the other with ambiguous ones (388 names).

In the previous experiment, two lists of male first names were used. The list of male first names provided with ANNIE contains 4,373 names. These names were merged with a list of the 1,216 most common male names in the 1990 US Census. These names occurred in 90% of the US population in 1990. This list was merged with the male first names from other lists including those of Presidential Nominees and Appointees, and US Senators and Representatives, and White House staff to create a list of male first names with 4,821 entries. This list of names was compared with city and dictionary term in the same manner as female first names, to generate a list of male unambiguous first names (3704 names) and a list of male first names that are ambiguous (1,117 names).

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 chiefs of state and heads of government was created from the 1990 CIA World Factbook [CIA 1990]. A list of the full names of ambassadors to the US was also extracted from the Factbook.

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	Unambiguous female first names from US 1990 census merged with ANNIE's default list and other lists	person_first	female	8,052
person_female_first_ambig.lst	Ambiguous female first names	person_first	female_ambig	388
person_female_first_ambig_cap.lst	Ambiguous female first names (in upper case)	person_first	female_ambig	388
person_female_first_cap.lst	Unambiguous female first names (in upper case)	person_first	female	8,052
person_headofstate_90.lst	Full names of heads of state and heads of government from the 1990 CIA World Factbook	person_full		478
person_headofstate_90_last.lst	Same as preceding, but only last names	person_last		450
person_headofstate_90_last_cap.lst	Same as preceding, but in upper case	person_last		450
person_male_first.lst	Unambiguous male first names from US census, merged with ANNIE's default list and	person_first	male	3,704

Wordlist	Description	Major Type	Minor Type	Count
	other lists			
person_male_first_ambig.lst	Ambiguous male first names	person_first	male_ ambig	1,117
person_male_first_ambig_cap.lst	Ambiguous male first names (in upper case)	person_first	male_ ambig	1,117
person_male_first_cap.lst	Unambiguous male first names (in upper case)	person_first	male	3,704
person_surname.lst	Unambiguous surnames from US census of 1990 merged with other lists	person_last		83,805
person_surname_ambig.lst	Ambiguous surnames	person_last	ambig	6,802
person_surname_ambig_cap.lst	Ambiguous surnames (in upper case	person_last	ambig	6,802
person_surname_cap.lst	Unambiguous surnames (in upper case)	person_last		83,805

Figure 1. Additional Wordlists Used for Recognizing Persons Names

3.2 Titles and Position Names

Job titles of Presidential appointees appear in Presidential records. A list of these titles was created from the announcements of Presidential Appointments published in the Federal Register and reprinted in the Bush Presidential Public Papers. Additional appointee titles were extracted from Appendix A of the Bush Public Papers.

The position names of Presidential nominees to Federal office that must be approved by the Senate also appear in Presidential records. A list of these was created from the announcements of nominations in the Bush Public papers and the list of Presidential Nominations in Appendix B of the same.

The titles of Chiefs of State and Heads of Government also appear in Presidential records. A list of these was extracted from the CIA World Factbook for 1990.

A list of job titles of White House staff is already being used. Additional White House Staff titles were identified from announcements of appointments in the Bush Public Papers and added to the list. Figure 2 summarizes the additions to word lists for job titles.

Wordlist	Description	Major Type	Minor Type	Count
jobtitle_bush41_appts.lst	Titles of Appointees from the Bush Public Papers 1989-93	jobtitle		76
jobtitle_bush41_nominees.lst	Nominees to Federal Office requiring Senate approval, 1989-1992	jobtitle		296

Wordlist	Description	Major Type	Minor Type	Count
jobtitle_bush41_nominees_cap.lst	Nominees to Federal Office requiring Senate approval, 1989-1992	jobtitle		296
jobtitle_bush41_wh_staff.lst	White House staff titles	jobtitle		213
jobtitle_bush41_wh_staff_cap.lst	Same as White House staff titles, but in all caps	jobtitle		213
jobtitle_foreign_headofstate_90.lst	From CIA World Factbook 1990	jobtitle		119
jobtitle_foreign_headofstate_90_cap.lst	From CIA World Factbook 1990	jobtitle		119

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

3.3 Locations and Facilities

In a 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 annotations were only partially correct, and there were six false positives. The missed location names were primarily US city and county names.

City and county names make up many of the location names in the Bush e-records. The list of city names provided with the vanilla information extractor (ANNIE) contains 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.

The Geographic Names Information System (GNIS) is the Federal standard for geographic nomenclature.[3] The Topical Gazetteer of Populated Places included in the GNIS contains named features with human habitation—cities, towns, villages, etc. It contains 183,771 entries. City, town and village names were extracted from this database and duplicates eliminated. Many US city names are ambiguous. They are also person's names (Alexander, Arkansas), the names of concepts (Independence, Missouri), or other objects (Alligator, Mississippi). Consequently, the list of US city names was divided into two lists, a list of unambiguous names with 33,017 entries and a list of ambiguous city names with 5,478 entries.

The Topographic Gazetteer of Populated Places also includes US county names. The county names were extracted and a wordlist created.

The United Nations *Demographic Yearbook* reports the population of major international cities as well as the number of births, deaths, marriages and divorces in those cities. The population of capital cities and cities of 100,000 and more inhabitants was downloaded in

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

Excel spreadsheet format from the Demographic Yearbook 2004.[4] US City names were deleted from this spreadsheet leaving 3770 foreign city names. There were some duplicate city names such as Colombo, the name of a large city in Sri Lanka as well as Brazil. Duplicates were eliminated. There were a number of cities for which English language alternative spellings of foreign city names were added, for example, Venice for Venezia and Rome for Roma. Some of the foreign city names were ambiguous, for example, "Nice, France" and "Nice day".

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.[5] In the future, this database could be used to create a more complete list of foreign city names.

In a previous experiment, "World Congress Center," "CNN Center" and "Omni Hotel" were recognized as entity type organization, when in the documents they are actually the names of facilities. A wordlist called loc_facility_post.lst was created that contains words such as Hotel, Airport and Center. JAPE rules were created that recognize proper nouns followed by one of these terms as the name of a location of type facility.

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

Wordlist	Description	Major Type	Minor Type	Count
loc_city_us.lst	US City, town and village names, except those that are ambiguous.	location	city_us	33,017
loc_city_us_ambig.lst	Ambiguous US City, town and village names that are also the names of states, persons and concepts such as independence and constitution	location	city_us_ambig	5,478
loc_city_us_ambig_cap.lst	Same as loc_city_us_ambig.lst but in all caps	location	city_us_ambig	5,478
loc_city_us_cap.lst	Same loc_city_us.lst, but in all caps	location	city_us	33,017
loc_facility_post.lst	Suffixes of facility names such as Hotel, Center, Airport and Air Force Base	location	facility_post	24
loc_foreign_city.lst	Names of capitals of foreign countries and cities with population of 100,000 or more.	location	city_foreign	3,802
loc_foreign_city_ambig.lst	Ambiguous names of	location	city_foreign_ambig	100

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

Wordlist	Description	Major Type	Minor Type	Count
	capitals of foreign countries			
loc_foreign_city_ambig_cap.lst	Ambiguous names of capitals of foreign countries (in all caps)	location	city_foreign_ambig	100
loc_foreign_city_cap.lst	Same as preceding list, but in upper case.	location	city_foreign	3,802
loc_us_county.lst	US counties and parishes	location	county	1938
loc_us_county_cap.lst	Same as preceding list, but in upper case.	location	county	1948
loc_us_state_abbr_ambig.lst	State abbreviations that are ambiguous such as IN and ME	location	state_ambig	9

Figure 3. Additional Wordlists for Locations and Facilities

3.4 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 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,[6] 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 US Government Departments and Agencies was found. Abbreviations for the departments and agencies were also entered in a wordlist.

Names of foreign governments often differ from the name of the country, for example, the French Republic and France. Names of governments were extracted from the CIA World Factbook. Names of governments are annotated as organizations of minor type government.

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 US colleges and universities was created.

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

Figure 4 summarizes additions to organization name wordlists.

Wordlist	Description	Major Type	Minor Type	Count
org_broadcast_media.lst	Names of TV, radio and cable news organizations such as NBC News	organization	news_ media	11
org_college_university.lst	Names of US colleges and universities.	organization	educational	1754
org_college_university_cap.lst	Same as preceding, but all caps	organization	educational	1754
org_government.lst	Names of foreign governments from CIA Factbook 1990	organization	government	195
org_media_company.lst	Names of companies that own newspapers and other media organizations	organization	news_ media	3
org_news_agency.lst	Names of wire services such as AP and UPI	organization	news_ media	5
org_news_magazine.lst	Names of US news magazines such as Newsweek and Time	organization	news_ media	7
org_newspaper.lst	Names of US newspapers	organization	news_ media	2829
org_newspaper_cap.lst	Same as preceding but capitalized	organization	news_ media	2829
org_us_govt_dept_agency.lst	Names of US government departments and agencies.	organization	government	519
org_us_govt_dept_agency_abbr.lst	Abbreviations of US government departments and agencies	organization	government	130
org_us_govt_dept_agency_cap.lst	Same as US government departments and agencies, but in all caps.	organization	government	519
org_us_hospitals.lst	Names of US hospitals	organization	hospital	3892
org_us_hospitals_cap.lst	Same as US hospitals, but all caps	organization	hospital	3892

Figure 4. Additions to Organization Name Wordlists

3.5 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.

3.6 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	8 years
second day of October	last few days
1976	4 days of this week
4-year Limitation	weekend
2/28	this weekend
six months ago	two weeks ago
WITHIN 9 WORKING DAYS	present
midnight EDT on April 25	

Relative temporal expressions, such as "Day after tomorrow," need to be recognized in order to understand sequences of actions or events mentioned in the text. To recognize relative temporal expressions, the wordlists were expanded beyond date units such as day, week, and hour into adverbs and adjectives, such as last, couple, early, and late, that modify these time units. A basic time modifiers list (time_mods2.lst) was created that includes words such as the following:

few	current	now
last	past	early
this	couple	late

Time expressions also use *frequency* words such as often, only, once, and twice. Frequency terms indicate the number of occurrences of an event in a time period. Consider the example: “We met just once in the past couple of days”. In the earlier implementation, 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	more	once
further	less	twice
just	most	sometimes
only	nearly	

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

usually	continuously	hardly
periodically	endlessly	barely
repeatedly	always	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	only
further	more
just	

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

endless	rare
periodic	bare
recurrent	scarce
steady	

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	frequency	adj	13
adjFreqPrefix.lst	Frequency adjectives	frequency	adjPrefix	2
advFreq.lst	Frequency adverbs	frequency	adv	17
advFreqDegree.lst	Degree of Frequency adverbs	frequency	advDegree	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

3.7 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
abbreviations_bill.lst	Abbreviations for types of congressional bills, e.g., H.R., S.	bill	type	29

4. Improvements to JAPE Rules

4.1 Rules Added to Correct Annotation Errors

4.1.1 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 "President", “General”, or “Chief of Staff”, the job title is considered to be the name of a person. The following JAPE rule produces such annotations.

```
Phase:  JobtitlePossPerson
Input: Lookup Token JobTitle TempOrganization
Options: control = appelt

Rule: JobtitlePossPerson1
(
 ({Token.string == "The"} | {Token.string == "THE"})
 {JobTitle}
)
:jobtitle
-->
 :jobtitle.Possperson = {rule = "JobTitlePossPerson1"}
```

4.1.2 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.

4.1.3 Names of Law Firms

A name of a law firms is 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
(

  (((LASTNAME) | {Token.orth == upperInitial})
  {Token.string == ","}
  )+
    ((LASTNAME) | {Token.orth == upperInitial})
    {Token.string == "&"}
      ((LASTNAME) | {Token.orth == upperInitial})

)
:orgName -->
    :orgName.TempOrganization = {kind = "lawfirm", rule = "LawFirm"}
```

4.1.4 Social Security Numbers

JAPE rules were added to recognize social security numbers in text. A macro was created to recognize words that precede social security numbers such as: Social Security Number, SSN#, SSN#: and SSN. Two rules were created to search for social security numbers in the format of: 444-44-4444 or 9 digit number sequences. An example rule is shown below.

```
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.SSN = {kind = "fullNumber", rule = "SSN_Number"}
```

4.1.5 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
// million bucks
// penny
// $30 million

(
 ({Token.symbolkind == currency})?
 (AMOUNT_NUMBER)?
 {Lookup.majorType == currency_unit, Lookup.minorType == post_amount}

)
:number

 -->
  :number.Money = {kind = "number", rule = "MoneyUnit"}
```

4.1.6 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 is included in the address annotation when the recipient is an organization. Following is an example of a rule that annotates one type of military address:

```
 Rule: StreetMilitary
// us mod:
(
(({Token.string == "The"})?
  ({TempOrganization} | {Lookup.majorType == organization})
)?
 ({Token.string == "PSC"}
  {Token.kind == number})?
 ({Token.string == "APO"} | {Token.string == "FPO"})
 ({Token.string == "AE"} | {Token.string == "AP"} | {Token.string == "AA"})
 {Token.kind == number}
 ({Token.string == "-"}
  {Token.kind == number})?
  ({Lookup.minorType == city_US_ambig} | {Lookup.minorType == city_US} )
  ({Token.string == ","})?
  ({Lookup.minorType == state_ambig} | {Lookup.minorType == state})
 (ZIP_CODE)?
):streetAddress -->
 :streetAddress.Address = {king = "streetAddress", rule = "StreetMilitary"}

 /*
```

The corpus includes military addresses, civilian addresses, addresses of convenience (The White House, Washington, DC) and international addresses. Rules were written and modified to adjust for annotation of the whole address and the various ways in which the address form can come. It was important to captures the different address types in their myriad forms as the address annotation is used in document type recognition.

4.2 Rules for Relative Temporal Expressions

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

In the second table, the pattern accounts for combined date/date words optionally separated a time expansion expression such as "to the" or "to".

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

The following rule is used to recognize the temporal expressions in the third 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"}
```

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

4.3 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."

"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."

"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."

"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."

"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)."

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." [7]

A *public law* is "A public bill or joint resolution that has passed both chambers and been enacted into law." [8]

Public laws and sections and titles of the United States Code are also mentioned in Presidential records, for example,

"… including the Comprehensive Anti-Apartheid Act of 1986 (Public Law 99 - 440), as amended (``the Act''), and section 301 of title 3 of the United States Code …"

The following JAPE rule takes bill and statute prefixes like HR, S and S.Res and looks for following multiple digits, for example, 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 foll
(
{Lookup.majorType == bill, Lookup.minorType == prefix}
({Token.kind == number})+
)
:bill --> :bill.Legislation = {kind = "billName", rule = BillPrefix}
```

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

JAPE rules have been created to annotate the names of acts. One rule matches a sequence of initially capitalized words preceding the word Act. Another JAPE rule recognizes *lists* of initially capitalized words separated by commas preceding the word Act and annotates them as an Act, for example, Financial Institutions Reform, Recovery, and Enforcement Act of 1989.

5. Tests to Verify Improvements in Semantic Annotation

5.1 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 of 50 documents are shown below [Isbell et al 2006].

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. 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 in Figure 6.

Annotation Type	Correct	Partially Correct	Missing	Spurious	Precision	Recall	F-Measure
Person	367	18	37	11	0.9495	0.891	0.9193
Location	339	16	28	7	0.9586	0.906	0.9315
Organization	395	28	20	13	0.9381	0.9233	0.9306
Date	281	3	2	3	0.9843	0.9878	0.986
Money	22	1	0	1	0.9375	0.9783	0.9574
Percent	28	0	0	0	1.0	1.0	1.0

Overall average precision: 0.9473 Overall average recall: 0.9297 Overall average F-measure: 0.9326

Figure 6. Results of a Test of the Performance of the Semantic Tagger on Corpus 2

In this particular test, average precision rose from 0.858 to 0.947, average recall from 0.806 to 0.929, and average F-measure from 0.832 to 0.932.

5.2 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. Figure 7 show excerpts from the results of the test. All names and titles in the test corpus were correctly recognized and annotated. Text phrases highlighted in purple are job titles, in brown are person's names, and in green are locations of subtype country.

Figure 7. Test of Annotation of Names of Chiefs of State/Heads of Government and their Titles.

5.3 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 8 shows an excerpt from the results of the test. All names and titles in the test corpus were correctly recognized and annotated.

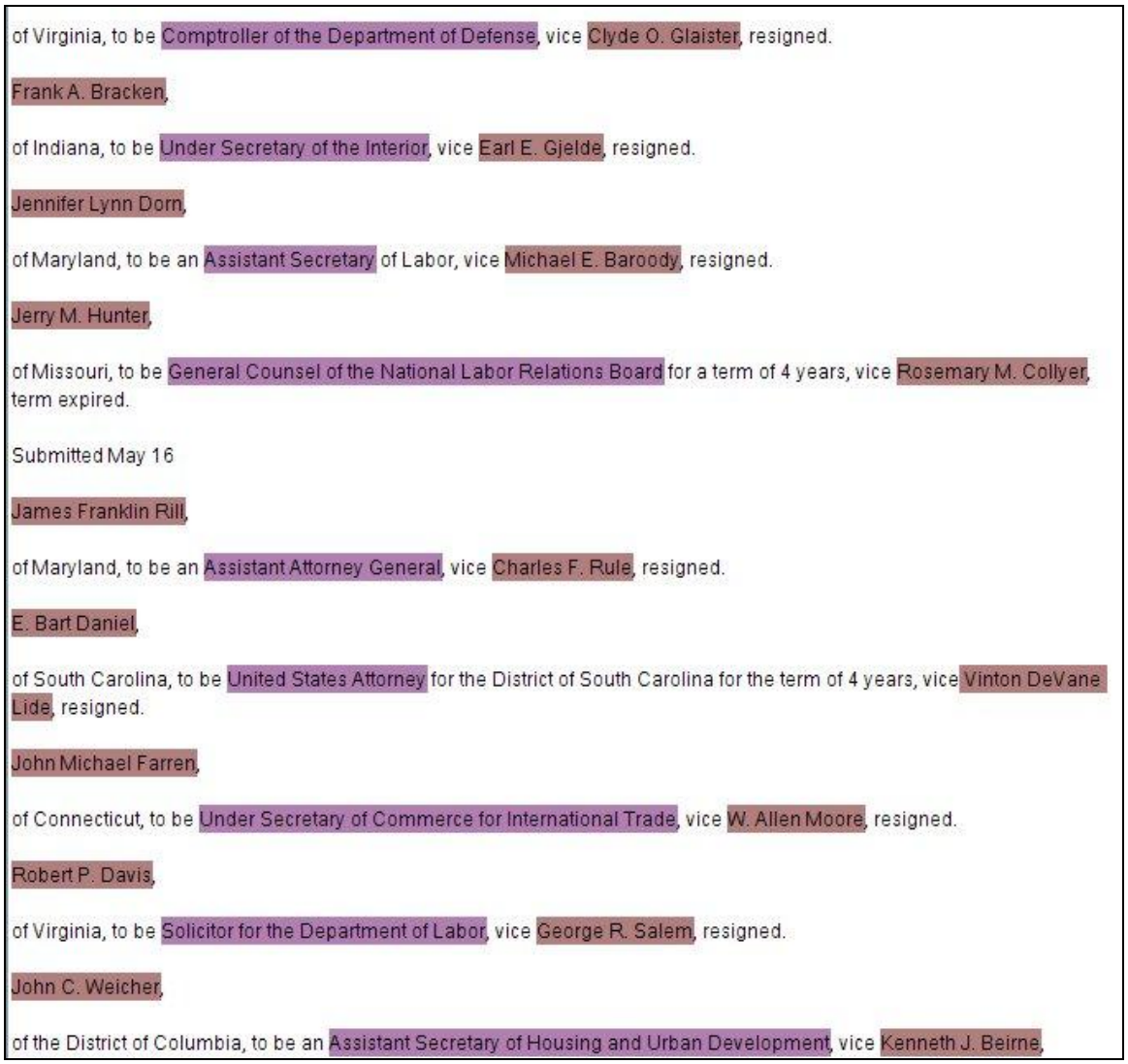

Figure 8. Excerpt from the Test of the Annotation of Names and Titles of Presidential Nominees

5.4 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 recognition and annotation of the names of bills and statutes. Figure 9 shows an excerpt from the test of the annotation of Congressional bills and statutes using the new JAPE rules. All bills and statutes in the test corpus were correctly recognized and annotated.

H.R. 840 / Public Law 101 - 92

To authorize appropriations for fiscal year 1990 for the Federal Maritime Commission, and for other purposes

H.R. 1426 / Public Law 101 - 93

Drug Abuse Treatment Technical Corrections Act of 1989

H.R. 2727 / Public Law 101 - 94

Court of Veterans Appeals Judges Retirement Act

Approved September 13

S.J. Res. 109 / Public Law 101 - 95

To designate the period commencing September 11, 1989, and ending on September 15, 1989, as ``National Historically Black Colleges Week''

Approved September 15

S.J. Res. 132 / Public Law 101 - 96

Designating September 1 through 30, 1989 as ``National Alcohol and Drug Treatment Month''

Approved September 23

H.R. 2136 / Public Law 101 - 97

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 Services Week''

S. 1075 / Public Law 101 - 99

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

5.5 Evaluation of the Semantic Annotation of Additional Semantic Categories.

Semantic annotation of job titles, postal addresses, legislation, and phone numbers was evaluated using corpus 2. The results are shown below.

Annotation Type	Correct	Partially Correct	Missing	Spurious	Precision	Recall	F-Measure
JobTitle	283	14	3	16	0.92651	0.96666	0.9461
Address	13	1	1	0	0.96428	0.9	0.9310
Legislation	19	2	2	1	0.90909	0.86956	0.8888
Telephone	8	0	0	0	1.0	1.0	1.0
SSN	0	0	0	0	0.0	0.0	0.0

Overall average precision: 0.9112 Overall average recall: 0.9238 Overall average F-Measure: 0.9034

Figure 10. Test of Annotation of Job Titles, Addresses, Legislation and Phone Numbers

The one missing address was:

1201 East Colfax Suite 220
Denver, Colorado 80218

This missed annotation is due to the fact that East Colfax is also a city. The address rule can be modified to find these kinds of missed annotations.

One of the missing legislation annotations, “Federal Home Loan Bank Act”, was due to “Federal Home Loan Bank” being annotated as an organization. A rule can be written to look for organizations followed by a legislative marker such as “Act”.

Many of the other missed and partial annotations for Legislation, JobTitle, and Address had similar errors that would require a modification or addition of a rule or two.

All telephone numbers were correctly annotated. There were not any social security numbers in corpus 2.

6. Experiment in Semantic Annotation of Presidential E-Records

Tools for extracting files from containers, for converting the files to text formats, and for annotating their contents are installed on the PERPOS system in the Virtual Laboratory at Archives II. They were used in an experiment to evaluate the performance of the semantic annotation of actual, born digital, personal computer records from the Bush Administration.

A corpus of 50 records was selected from a record series accessioned into PERPOS. They are records from the Office of Legislative Affairs. The Office of Legislative Affairs provides advice and support regarding the President's legislative agenda and legislation in general, and liaison between the White House staff and members of Congress.

The experiment evaluated 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 results are shown in Figure 11.

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

Overall average precision: 0.9178 Overall average recall: 0.9282 Overall average F-measure: 0.9108

Figure 11. The Performance of Semantic Annotation of Corpus 3.

The eight examples that were spuriously annotated as money were all instances of "mark up" of legislation. The term "mark" was annotated as Money (German Mark). This error is easily repaired by a rule that differentiates "mark up" of legislation from the monetary term "Mark".

Most of the missing annotations for locations were state abbreviations (OK, MA, ME, IN, VA) that appeared in parentheses after a legislator's name, indicating the state they represented. These missing location abbreviations were in a list of ambiguous state abbreviations. The ambiguous state abbreviations were only annotated in conjunction with an city name. However, they can easily be disambiguated in this case since they appear after a person's name and are in parentheses.

The table in Figure 12 shows:

1. The performance on Corpus 1 of the "vanilla" Semantic Annotator provided with the GATE distribution [Underwood 2004]
2. The performance on Corpus 2 of the Semantic Annotator after improvements made based on an analysis of Experiment 1 [Isbell 2006]
3. The performance on Corpus 3 of the Semantic Annotator after improvements to the wordlists and JAPE rules as described in this report.

	Experiment 1: Default ANNIE Corpus 1	Experiment 2: GaTech IE Vers 1 Corpus 2	Experiment 3: GaTech IE Vers 2 Corpus 3
Person	0.6768	0.7269	0.9044
Location	0.7926	0.7922	0.8565
Organization	0.6342	0.7182	0.9035
Date	0.8934	0.9471	0.9967
Money	1.0000	0.9286	0.8636
Percent	0.8182	0.9273	1.0000
Overall Average F-measure	0.7490	0.8316	0.9108

Figure 12. Improvements in F-measure in Three Experiments

Figure 13 shows in graphical form the improvements in performance.

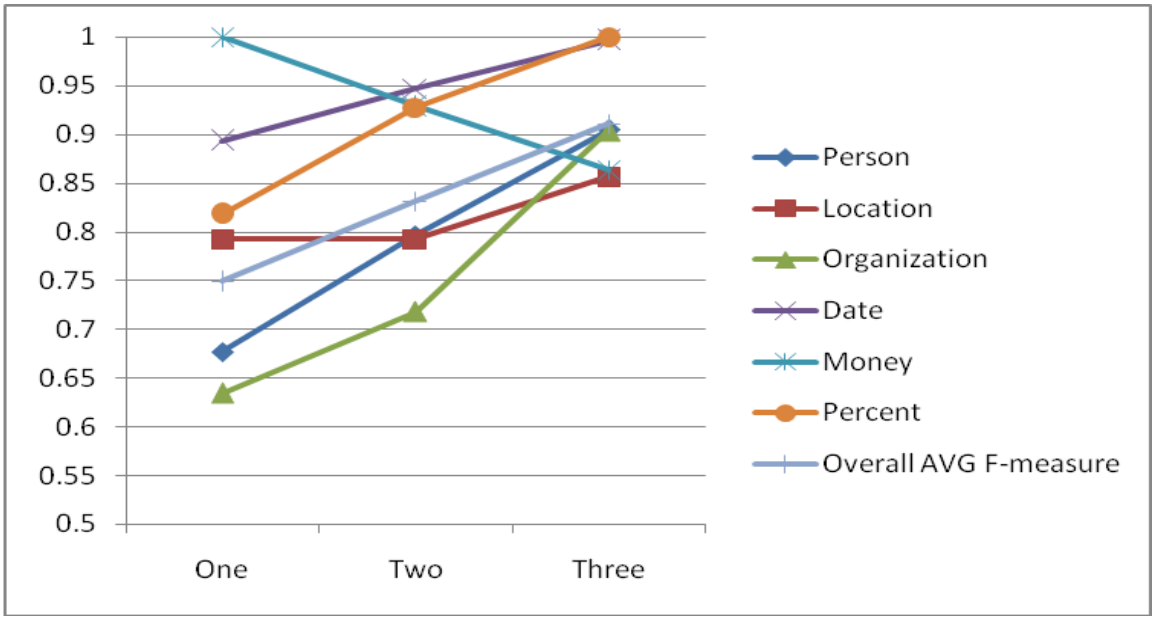

Figure 13. Graph Showing Improvements in the Performance of the Semantic Annotator

In the three experiments, the performance has increased in all cases with the exception of the annotation of money terms in experiment 3. As previously explained, in experiment 3, the decrease in performance in annotation of money terms is due to the incorrect annotation of "mark" in the "mark up" of legislation.

The current performance is very good. Without this level of performance, methods for speech act, topic and document type recognition, which are dependent on the semantic annotation method, cannot achieve a high-level of performance. However modifications will be made to correct the annotations in Experiment 3 that were missed, partially correct and spurious.

7. Summary of Results

The results of a previous information extraction experiment were analyzed to determine reasons for partially correct annotations, missing annotations and false positives. The solutions to the problems were in the provision of additional wordlists for semantic categories, disambiguation of current lists, and additional or modified JAPE rules. Additional wordlists and JAPE rules were created for recognizing and annotating new semantic categories, namely, relative temporal expressions, job titles, postal addresses, names of legislation and social security numbers.

Previously, experiments were conducted on copies of paper Presidential records that were scanned and OCRed. The current experiment used e-records from the Bush Presidential personal computer records. An experiment was conducted to evaluate the performance of the semantic annotator with regard to the named entities addressed in the previous two experiments, namely, names of persons, locations, organizations, dates, money and percents. The experiment showed significant improvements in performance for the annotation of each of these categories with the exception of money, and the annotation errors for that category are easily fixed. The level of performance achieved is adequate to support the development of methods for document type, speech act and topic recognition.

References

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

[Cunningham et al 2007] H. Cunningham , D. Maynard, K. Bontcheva, V. Tablan, C. Ursu, M. Dimitrov, M. Dowman, N. Aswani, I. Roberts, Y. Li, A. Shafirin. Developing Language Processing Components with GATE Version 4 (a User Guide). The University of Sheffield, 2007

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

[Harris and Underwood 2004] B. Harris and M. Underwood. Factual Knowledge Needed for Information Extraction and FOIA Review, PERPOS Technical Report 04-7, December, 2004.

[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 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

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

[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, Georgia Tech Research Institute, November 2006.

[Underwood 2004] M. G. Underwood. Recognizing Named Entities in Presidential Electronic Records**,** PERPOS Technical Report ITTL/CISTD 04-4, Georgia Tech Research Institute, June, 2004 (Revised Nov 2004).

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

Appendix A: Wordlists

File Name	Major Type	Minor Type	Count
abbreviations.lst	stop		4
abbreviations_bill.lst	bill	prefix	29
addr_directional.lst	address	directional	12
addr_directional2.lst	directionals	all	22
addr_secondary_unit.lst	address	secondary_designation	16
addr_secondary_unit_cap.lst	address	secondary_designation	16
addr_secondary_unit_rng.lst	address	secondary_designation_rng	24
addr_secondary_unit_rng_cap.lst	address	secondary_designation_rng	24
addr_street.lst	address	street	9
addr_street_directional.lst	address	directional	8
addr_street_directional_abbr.lst	address	directional	12
addr_street_directional_cap.lst	address	directional	8
addr_streetspan.lst	address	street	33
addr_streetspan_abbr.lst	address	street	29
addr_streetspan_cap.lst	address	street	33
adj_country.lst	country_adj		782
adj_country2.lst	country_adj		10
adjFreq.lst	frequency	adj	13
adjFreqPrefix.lst	frequency	adjPrefix	2
advFreq.lst	frequency	adv	17
advFreqDegree.lst	frequency	advDegree	10
appendixFTitles.lst	jobtitle		4
currency_prefix.lst	currency_unit	pre_amount	11
currency_unit.lst	currency_unit	post_amount	258
date_day.lst	date	day	27
date_day_cap.lst	date	day	27
date_festival.lst	date	festival	156
date_key.lst	date_key		9
date_month.lst	date	month	114
date_ordinal.lst	date	ordinal	70
date_ordinal_cap.lst	date	ordinal	41
date_post.lst	date	post	52
date_unit.lst	date_unit		27
date_us_fed_holiday.lst	date	festival	14
date_us_fed_holiday_cap.lst	date	festival	14
date_year.lst	year		58
ident_prekey.lst	ident_key	pre	13
jobtitle.lst	jobtitle	TC	75
jobtitle_bush41_appts.lst	jobtitle		76
jobtitle_bush41_nominees.lst	jobtitle		296
jobtitle_bush41_wh_staff.lst	jobtitle		213
jobtitle_bush41_wh_staff_cap.lst	jobtitle		213
jobtitle_cap.lst	jobtitle	UC	75
jobtitle_foreign_headofstate_90.lst	jobtitle		119
jobtitle_modifiers.lst	jobtitle_modifiers		17
jobtitle_modifiers_cap.lst	jobtitle_modifiers		17
jobtitle_nom_appr.lst	jobtitle		50

jobtitle_nom_appr_cap.lst	jobtitle		50
loc_cities_us.lst	location	city_US	33,017
loc_cities_us_ambig.lst	location	city_US_ambig	5,478
loc_cities_us_ambig_cap.lst	location	city_US_ambig	5,478
loc_cities_us_cap.lst	location	city_US	33,017
loc_country.lst	location	country	458
loc_country_cap.lst	location	country	458
loc_facility_post.lst	location	facility_post	24
loc_facility_post_cap.lst	location	facility_post	24
loc_foreign_city.lst	location	city_foreign	3,802
loc_foreign_city_ambig.lst	location	city_foreign_ambig	100
loc_foreign_city_ambig_cap.lst	location	city_foreign_ambig	100
loc_foreign_city_cap.lst	location	city_foreign	3,802
loc_generalkey.lst	loc_general_key		10
loc_key.lst	loc_key	post	54
loc_mountain.lst	location	region	5
loc_prekey.lst	loc_key	pre	29
loc_prekey_lower.lst	loc_key	pre	29
loc_region.lst	location	region	67
loc_region_cap.lst	location	region	67
loc_us_county.lst	location	county	1,938
loc_us_county_cap.lst	location	county	1,948
loc_us_state.lst	location	state	50
loc_us_state_abbr.lst	location	state	44
loc_us_state_abbr_ambig.lst	location	state_ambig	9
loc_us_state_ambig.lst	location	state_ambig	3
loc_us_state_ambig_cap.lst	location	state_ambig	3
loc_us_state_cap.lst	location	state	50
loc_water.lst	location	region	159
loc_water_cap.lst	location	region	159
numbers.lst	number		59
org_base.lst	org_base		92
org_base_cap.lst	org_base		97
org_broadcast_media.lst	organization	news_media	11
org_charities.lst	organization	charity	79
org_college_univ.lst	organization	educational	1,754
org_college_univ_cap.lst	organization	educational	1,754
org_company.lst	organization	company	2,607
org_company_cap.lst	organization	company	2,562
org_company_designator.lst	cdg		142
org_congressional_cmte.lst	organization		67
org_congressional_cmte_cap.lst	organization		67
org_ending.lst	org_ending		138
org_foreign.lst	organization		61
org_foreign_govt_dept_agency.lst	organization	government	32
org_govern_key.lst	govern_key		23
org_govern_pre.lst	org_pre		5
org_govern_pre_cap.lst	org_pre		5
org_government.lst	organization	government	195
org_judicial.lst	organization		33
org_judicial_cap.lst	organization		33
org_key.lst	org_key		110
org_key_cap.lst	org_key	cap	110

org_media_company.lst	organization	news_media	3
org_ministry.lst	organization	government	16
org_news_agency.lst	organization	news_media	5
org_news_magazine.lst	organization	news_media	7
org_newspaper.lst	organization	news_media	2,829
org_newspaper_cap.lst	organization	news_media	2,829
org_noun.lst	organization_noun	upper	915
org_noun_cap.lst	organization_noun		915
org_noun_lower.lst	organization_noun		915
org_patriotic_frat_civic.lst	organization		23
org_patriotic_frat_civic_cap.lst	organization		23
org_pre.lst	org_pre		19
org_pre_cap.lst	org_pre		18
org_public_policy.lst	organization	public_policy	1,224
org_public_policy_cap.lst	organization	public_policy	1,224
org_spur.lst	spur		4
org_suffix_new.lst	cdg		259
org_us_govt_dept_agency.lst	organization	government	519
org_us_govt_dept_agency_abbr.lst	organization	government	130
org_us_govt_dept_agency_cap.lst	organization	government	519
org_us_hospitals.lst	organization	hospital	3,892
org_us_hospitals_cap.lst	organization	hospital	3,892
org_us_industry_trade_assoc.lst	organization	trade_assoc	42
org_us_industry_trade_assoc_cap.lst	organization	trade_assoc	42
org_us_political_party.lst	organization		5
org_us_political_party_cap.lst	organization		5
org_wh_office.lst	organization		77
org_wh_office_cap.lst	organization		77
person_ambassador_to_us.lst	person_full	normal	148
person_ambassador_to_us_cap.lst	person_full	normal	148
person_ending.lst	person_ending		17
person_female_first.lst	person_first	female	8,052
person_female_first_ambig.lst	person_first	ambig	388
person_female_first_ambig_cap.lst	person_first	ambig	388
person_female_first_cap.lst	person_first	female	8,052
person_full.lst	person_full	normal	25
person_full_cap.lst	person_full	normal	25
person_head_of_state_last.lst	person_last		450
person_head_of_state_last_cap.lst	person_last		450
person_headofstate_90.lst	person_full	normal	478
person_headofstate_90_cap.lst	Person_full	normal	478
person_male_first.lst	person_first	male	3,704
person_male_first_ambig.lst	person_first	ambig	1,117
person_male_first_ambig_cap.lst	person_first	ambig	1,117
person_male_first_cap.lst	person_first	male	3,704
person_surname.lst	person_last		83,805
person_surname_ambig.lst	person_last	ambig	6,802
person_surname_ambig_cap.lst	person_last	ambig	6,802
person_surname_cap.lst	person_last		83,805
person_surname_prefix.lst	surname	prefix	11
person_unitednations.lst	person_full	normal	2
phone_prefix.lst	phone_prefix		14
spur_ident.lst	spur_ident		0

stop.lst	stop		28
time_ampm.lst	time	ampm	12
time_hour.lst	time	hour	13
time_key.lst	time_modifier		12
time_key_cap.lst	time_modifier		12
time_modifier.lst	time_modifier		10
time_modifier_cap.lst	time_modifier		5
time_mods2.lst	time_modifier		13
time_mods2_cap.lst	time_modifier		13
time_suffix.lst	time_suffix		1
time_unit.lst	time_unit		7
times.lst	time		39
timespan.lst	time_span		9
timespan_cap.lst	time_span		9
timex_pre.lst	time_modifier		26
timex_pre_cap.lst	time_modifier		26
timezone.lst	time	zone	69
timezone_cap.lst	time	zone	69
title.lst	title	civilian	169
title_cap.lst	title	civilian	169
title_female.lst	title	female	13
title_female_cap.lst	title	female	13
title_male.lst	title	male	10
title_male_cap.lst	title	male	10
title_mil.lst	title	military	259
title_mil_cap.lst	title	military	260
title_police.lst	title	police	72
title_police_cap.lst	title	police	72
Total			335,457