History Meets Computer Science
Intelligent Access to Historical Documents

Nancy Ide
Department of Computer Science
Vassar College

Collaboration

- Department of Computer Science, Vassar College (PI: Nancy Ide)
- Franklin and Eleanor Roosevelt Institute and Department of History, Marist College (PI: David Woolner)
- Funding from NSF ITR for a three year pilot project

The FDR/Pearl Harbor Project

- enhancement of materials drawn from the *Franklin D. Roosevelt Library and Digital Archives*
- image, sound, video and textual data
- encoding, annotation, and multi-modal linkage of a portion of the collection
- enhancement of a web-based interface that enables exploitation of state-of-the-art methods for search and retrieval

Major Activities

- development of a model for data in historical documents relevant for historical research
- instantiation using W3C standards : XML, Resource Definition Framework (RDF and RDF schemas), Ontology Web Language (OWL)
- Enhancement/development of automated means to identify and mark relevant entities
- exploration of the potential to automatically extract ontological information to enable sophisticated search and retrieval via inferencing

The Data

- Government correspondence and documents produced in the sixth months prior to and including December 7, 1941, the date of the Japanese attack on Pearl Harbor
- The nature of the data and the uses to which it will be put differ from usual projects in Automated Language Processing

- “Domain specific” -- but a different type of domain
- Users are historians looking at detailed information, some non-explicit, from a variety of views (military, strategic, diplomatic, economic, etc.) Time line is critical
- Richer set of entity types and relations than typical, should be applicable in IE in general
- Some of same interests (document topics, who said what when etc.) but also interested in e.g. attitude conveyed by language
  - E.g., documents written by the same person on the same dates addressed to different audiences may reveal very different attitudes and concerns in description of same events

Application of established methods should provide insight into their potential to treat a wider range of document types

Pilot Corpus

- Critical collection of 100 key documents leading up to the Japanese attack on Pearl Harbor
- Focus on strategic, diplomatic and economic aspects of U.S.- Japanese relations in the six months prior to the attack
- Text types:
  - Letters, memoranda of conversations, proposals, press releases, notes, telegrams
- stylistically varied
  - e.g., telegrams contain cryptic, unpunctuated phrasing

The Content

- texts document growing military and economic tensions between the United States and Japan over e.g. the Japanese incursion into China and the increasing likelihood of a military confrontation
- internal White House documents generated during this period critical to an understanding of the events and attitudes leading up to the American declaration of war

Aims

- provide “intelligent” search and access for historians of the Second World War
  - support the data with an ontology in the background
  - enable retrieval not only on the basis of specific names, dates, persons, etc., but also category and/or role, document style, etc.
  - identification and classification of events
  - ultimately, exploit inferencing capabilities to unearth information that is not explicit or obvious

Data Preparation

- documents drawn from originals held in the Franklin D. Roosevelt Presidential Library
- scanned, hand-validated, and encoded in XML format according to the specifications of the XML Corpus Encoding Standard (XCES)
  - full XCES-compliant header
  - RDF meta-data specifications according to Dublin Core categories

Document images available from the FDR Library Digital Archives at http://www.fdrlibrary.marist.edu

Processing

- Using the Univ. of Sheffield's GATE (General Architecture for Text Engineering) system to annotate the data
  - Allows defining annotation patterns for entity and event recognition using a powerful language for pattern specification
- Annotations provided in GATE:
  - Token, sentence, part of speech, NP chunking, VP chunking
  - entities : person names, dates, locations, job titles, etc.

Entity Recognition

- Two sources of information for automatic recognition
  - Gazetteer lists
  - Annotation pattern rules
- Structure for scalability
  - E.g, have a gazetteer list of first names and last names, rule for combining, rather than a list with full names
- Identify variant orthographic forms as instances of the same entity where applicable

Refinements to Entity Recognition

- Refinement of annotation patterns and lists
  - Fine-tuning patterns for person names, variants of same name
  - Adding rich set of region and location names to gazetteer lists (e.g. Manchuko)
  - Adding job titles to gazetteer lists plus rules for more complex titles (e.g., Ambassador General, Chief of the Bureau of Far East, Minister-Counselor of the Japanese Embassy)
- Additional entities
  - document, policy, agreement, and treaty names, military groups and operations
  - references to “situations” (the China problem, the Manchuria situation)

Classification of entities/ontology

- Developing a finer-grained classification for
  - Job titles (e.g., head of state, chief executive, various levels of government positions)
  - Geographical regions, sub-regions of importance for our domain
    - Southwestern Pacific (Australia, New Zealand), Southern French-Indochina, eastern Siberia
- Classification of locations by areas relevant to WWII
  - Pacific theatre, Atlantic theatre, European theatre
- Classification of countries/regions by alliance/strategic relevance
  - Alliance: Axis/allied power, neutral power
  - Strategic importance : naval port/base, conduit (Panama Canal, Burma Road)
  - Colonies, puppet states, occupied territories

Ontology creation

- Two methods:
  1. Enter by hand - easy for our confined domain
     - But we want scalability in order to apply to more of the documents in FDR Library (and others)
  2. Automatic learning
     - E.g. **Netherlands** East Indies, **French** Indochina
     - Just starting this work
- Later, apply inferencing
  - E.g. Britian a Pacific power by virtue of its possessions
- **Question**: where to draw the line between supplied/inferred information

Ontology

- Using RDF Schemas and OWL
- Where possible, extending OpenCYC/DAML
- Much of our information can be described by extending the upper ontologies for government, military organizations, and people related to organizations

Extension/refinement

- Our data demands substantial refinement and (occasionally) re-definition of OpenCYC/DAML categories
  - E.g. key-members (someone who “is, or often gives input to, the organization's leader and thus may substantially influence the decisions of the organization”)
    - Our data: government officials but also others (members of the Japanese Imperial Family, various American personalities (e.g., Fred Kent, a New York banker, and E. Stanley Jones, a Methodist Minister)
- Tricky cases: e.g., status of France as a geo-political entity
  - “Vichy France” is official government but governs only southeastern France; not an Axis power (“collaborative” relationship with the Germans) but not an Allied power
  - western and northern France is “occupied France”, governed by the Germans
  - “Free France”- Charles DeGaulle’s counter-government located in London

Document clustering

- Use agglomerative clustering algorithm to classify texts
  - All words : topic (economic, diplomatic, strategic)
  - Also clustering by verbs, names, locations, and others
  - Cluster by style/genre
    - Use Biber's software, analyzes style/genre using over 70 linguistic features
- Likely, different document clusters will be relevant for different purposes/queries

Event recognition

- Identify general event types
  - **historical events** referred to in the documents (e.g., “the award against Japan by the Hague tribunal in the Perpetual Leases matter,” “Chinese troop movements along the northern frontier of French Indo-China”)
  - **communicative events** represented by the documents themselves (letter to X from Y)
  - **communicative events** reported in the documents, primarily in the Memoranda of Conversation
  - **conjectured events**, reflecting assertions about possible actions or results (e.g., “if the United States should expect that Japan was to take off its hat to Chiang Kai-shek and propose to recognize him, Japan could not agree”) and their relation to actual future events
  - **historical background**, events not directly referred to in the documents (e.g. the U.S. oil embargo against Japan)

Event Recognition

- Identify specific event types
  - **Move (x from y to z)**
  - **Communicate (x by y to z)** sub-types: agreement, disapproval, conciliatory, promise, etc.
  - **Positive/negative act (x by y affecting z)** sub-types : military, economic; sub-sub-types: embargo, “recognize”, etc.
- **Much richer/more detailed set of events than in newspapers etc.**

Event detection

- Currently focusing on identification of **communicative events** reported in the documents
  - extracted all verbs from the corpus, grouped them on the basis of WordNet 2.0 synsets
  - assigned a frame category to each group using FrameNet
  - extracted groups associated with communication frames and sub-frames
  - added information not in FrameNet
    - E.g, distinguish lexical units described by the “Judgment-communication” frame for negative or positive valency (e.g., “acclaim” and “condemn” belong to the same frame

Representation

- Need to determine data model that
  - Captures (all) the right information
  - Is in a form that most easily enables retrieval
  - Is extensible to other documents
- Using models identified in DARPA ACE etc. as a starting point, but need to enhance as well as add additional types of information

Summary

- Very constrained domain and clear use for our data enable us to
  - Create of a far more complex and richer KB/ontology than usual in NLP
  - Explore generality of established methods for entity detection, ontology learning
  - Explore use of a rich ontology for inferencing to support historical research and retrieval in general
  - Explore viability of semantic web technology to support historical research
    - Freely available web interface to data (all in public domain)

Collaboration potential

- Build on the expertise of historians to know what is what, what is important information etc.
- Build on CS methods for representation, search, and access
- Took about 8 months to fully understand one another
  - Historians learn to see their data in entirely new ways
  - CS folks learn new and challenging areas to apply techniques

Contact Information

lde@cs.vassar.edu