Exploiting Semantic Web Technologies for Intelligent Access to Historical Documents

Nancy Ide[1] and David Woolner[2]

[1]Department of Computer Science, Vassar College, Poughkeepsie, NY 12604-0520 USA

[2]Department of History, Marist College, Poughkeepsie, NY 12601 USA

ide@cs.vassar.edu, dwoolner@feri.org

Abstract

The FDR/Pearl Harbor Project involves the enhancement of materials drawn from the *Franklin D. Roosevelt Library and Digital Archives,* which includes a range of image, sound, video and textual data. The project is undertaking the encoding, annotation, and multi-modal linkage of a portion of the collection, and enhancement of a web-based interface that enables exploitation of state-of-the-art methods for search and retrieval. We are currently developing a pilot project that includes 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, which has obvious historical, political, and general interest. The major activities in the project involve development of a model for historical documents and associated data and its instantiation using W3 standards, including XML, the Resource Definition Framework (RDF and RDF schemas), and the Ontology Web Language (OWL); development of automated means, or enhancement of existing software, to identify and mark relevant elements within these data; and exploration of the potential to automatically extract ontological information so as to enable sophisticated search and retrieval via inferencing.

Introduction

The FDR/Pearl Harbor Project involves the enhancement of materials drawn from the *Franklin D. Roosevelt Library and Digital Archives,* which includes a range of image, sound, video and textual data. The project is undertaking the encoding, annotation, and multi-modal linkage of a portion of the collection, and enhancement of a web-based interface that enables exploitation of state-of-the-art methods for search and retrieval. We are currently developing a pilot project which includes 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, which has obvious historical, political, and general interest.

The major activities in the project involve development of a model for historical documents and associated data and its instantiation using W3 standards, including XML, the Resource Definition Framework (RDF and RDF schemas), and the Ontology Web Language (OWL); development of automated means, or enhancement of existing software, to identify and mark relevant elements within these data; and exploration of the potential to automatically extract ontological information so as to enable sophisticated search and retrieval via inferencing.

In this paper we describe the overall project design and the methodologies for annotating data for a variety of linguistic features (part of speech and shallow syntax, various named entities, time, events, etc.); (semi-) automatic derivation of ontological relations from the data; RDF/OWL-based representations of ontological relations and extensions to existing ontologies (OpenCYC, DAML); and discussion of representation choices based on processing and user requirements.

The nature of our data and the uses to which it will be put differ considerably from projects in our field, and therefore the FDR/Pearl Harbor project should provide insight into the applicability of established methods for language analysis to a wider range of document types than has been previously explored in depth.

Corpus Content and Use

The texts in our corpus constitute a critical collection of 100 key documents leading up to the Japanese attack on Pearl Harbor. They focus on the strategic, diplomatic and economic aspects of U.S.- Japanese relations in the six months prior to the attack. Among these are letters to and from President Roosevelt and various high-level U.S. government officials (Secretary of State, Ambassador to Japan, etc.); memoranda of conversations, primarily between U.S. officials and representatives of the Japanese government; proposals by Roosevelt and other high level U.S. officials on how to handle the situation with Japan; press releases; notes; and telegrams. The documents vary considerably stylistically; the most aberrant style appears in the telegrams, which often contain cryptic, unpunctuated phrasing intended to reduce the size of the message.

The texts in the corpus document the growing military and economic tensions between the United States and Japan over such issues as the Japanese incursion into China and Indochina, and the increasing likelihood of a military confrontation between Japan and the U.S. Given recent allegations that Roosevelt orchestrated the conflict with Japan in order to justify entering into the war against Germany, and theories that the Japanese intentionally misled the American government in the weeks leading up to the attack on Pearl Harbor, internal White House documents generated during this period are critical to an

understanding of the events leading up to the American declaration of war.

Our aim in the FDR/Pearl Harbor project is to provide "intelligent" search and access methods for historians of the Second World War. By supporting the data with an ontology in the background, we can enable retrieval not only on the basis of specific names, dates, persons, etc., but also by category and/or role (e.g., the Axis powers, advisors to the President, strategic naval ports, communications sent to Secretary Hull in October 1941). Events will also be represented and classified, allowing for retrieval of information such as "all assertions made (or all questions asked) by Ambassador Kurusu in conversations with U.S. government representatives between July and December 1941, in time-linear order". Ultimately, we intend to exploit inferencing capabilities that can unearth information that may not be explicit or obvious: for example, if we know that Roosevelt signed a document on a given date from Hyde Park, it can be inferred that he was physically present in Hyde Park on that date. This is a simple example, but one can imagine comparable inferences that will provide historians with unprecedented means to track the deterioration in US-Japanese economic relations as a parallel issue in relation to what was happening in the diplomatic or strategic sphere, and vice versa. If, for example, we retrieve documents written by the same person on the same dates, but which are addressed to different audiences (in particular, others within the same government vs. "public" diplomatic documents that will be read by members of other governments), a comparison of the events represented in each may reveal very different attitudes and concerns.

Data Preparation

The documents in the FDR pilot corpus are drawn from originals held in the Franklin D. Roosevelt Presidential Library.[1] The documents were scanned, hand-validated, and encoded in XML format according to the specifications of the XML Corpus Encoding Standard (XCES) (Ide, *et al*., 2000). Each documents includes a full XCES-compliant header as well as RDF meta-data specifications according to Dublin Core categories.

The FDR/Pearl Harbor Project is using the MUSE application[2] within the GATE (General Architecture for Text Engineering) system developed at the University of Sheffield (Cunningham, 2002) to annotate the data. The ability of various GATE components to enable definition of annotation patterns and to adapt to specific text types is obviously well-suited to our needs for entity and event recognition. GATE was used to annotate the data for part of speech, NP chunking, and VP chunking. We have also exploited GATE (and the MUSE application in particular) to identify and mark various entities such as person names, dates, locations, and job titles. Lemmas were added to the part of speech annotation in a post-processing step based on dictionary lookup, using the Multext lexicon for English.[3]

Our first step was to attempt to classify the texts by topic, using an agglomerative clustering algorithm provided in the Cluto software package (Zhao & Karypis, 2001). The texts were represented using a vector-space model, in which each document is represented by a term-frequency vector whose values are weighted based on inverse document frequency. Several terms were eliminated from the analysis because of their high frequency in all of the documents; these included (among others) "Japan", "America", "Ambassador", and "Secretary". Experimentation with varying numbers of clusters (5 –20) and terms (10 – 100) yielded fairly consistent results that partitioned the documents into two groups judged to be meaningful by a Roosevelt historian: one corresponding to documents concerned with economic matters, and another focusing on diplmatic/strategic concerns. Sub-clusters identified within these two groups were not judged to be topically distinguishable. We are continuing to experiment with different term sets and weighting mechanisms to determine if further topic categorization is possible in order to provide historians with more precise categories for text selection.

To identify named entities, the MUSE system operates in two fundamental steps: an "orthomatcher" that consults a gazetteer containing pre-defined lists of strings and tags those it matches according to specified categories, followed by rule-based entity recognition. Without enhancement and using only the lists included with the system, MUSE successfully identified about half of the name types in our corpus, missing primarily names preceded by titles specific to our corpus such as "Foreign Minister Ribbentrop" and "Ambassador General Oshima". Approximately 10% of the identified entities were erroneous; for example, MUSE returned several capitalized words presumably not in the gazeteer (e.g., "Spring", "Summer", "Inasmuch", and "Yen"-–but not "Hitler" and "Stalin"). Names specific to our corpus and their variants have been added to the MUSE gazetteer lists to yield almost 100% precision, although false hits must be edited out by hand. Similarly, we have augmented the MUSE gazetteer to include the large number of location and region names in our corpus that were not previously included in the MUSE lists, together with a variety of document, policy, agreement, and treaty names, military groups and operations, etc. In a post-processing step, linguistic information is exploited to resolve ambiguities, for example, distinction of "Japanese" as a noun (therefore referring to the Japanese people or government), from its use as an adjective.

One of the major challenges of this project is to appropriately represent entities in the data so that they are relevant for historical and political research. This often involves detailed analysis of internal structure in order to identify and mark relevant components: for example, an entity such as "Roosevelt Administration" requires both recognizing the entire string as a named entity, and marking "Roosevelt" as a name within a name. However, the greatest challenges for adequate representation of

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

[2] See http://gate.ac.uk/sale/muse/muse.pdf

[3] Because the POS tags provided by GATE and the Multext lexicon POS annotations differ, it was necessary to map the two for this step.

information in the FDR data are even more complex if we intend to provide sophisticated retrieval capabilities. For example, when dealing with names such as "Roosevelt Administration," it is necessary to address questions such as: What type of entity is "Roosevelt Administration"? What is the semantic relation of the person-name "Franklin Roosevelt" to the name "Roosevelt" in "Roosevelt Administration," and how do we represent it? If a scholar is searching for Roosevelt (the person), should "Roosevelt Administration" be retrieved or not? If we know that Roosevelt is the name of a person, can we infer that Roosevelt is the author of a document entitled "The Roosevelt Doctrine", and if so, should this information be made explicit or left to be determined on demand via on-the-fly inferencing? An important aspect of our work is to identify the information that can/should be represented in markup, information that can or must be represented as a set of ontological relations among objects, and information that can be inferred on-the-fly. We see this question as fundamental to data representation and retrieval on the Semantic Web: there is a trade-off between the effort required to mark the data and the processing overhead required to determine this information, dynamically or otherwise, that has received very little consideration before now.

Event Recognition

For historical research, identification of a range of "events" in our data is essential. Unlike many of the document types to which event recognition strategies have been applied (e.g., newswire), our data require recognition of several different categories of events: (1) historical events referred to in the documents (e.g., "the award against Japan by the Hague tribunal in the Perpetual Leases matter"); (2) communicative events represented by the documents themselves; (3) communicative events reported in the documents, primarily in the Memoranda of Conversation; and (4) 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"). Furthermore, the documents themselves make sense only against the backdrop of the series of well-known historical events that occurred during the six months before the war, such as the U.S. oil embargo against Japan, which may or may not be directly referred to.

As a first step, we are focusing on identification of communicative events reported in the documents. To accomplish this, we first extracted all verbs from the corpus and grouped them on the basis of WordNet 2.0 synsets[4]. We then assigned a frame category to each group by consulting the FrameNet database (Fillmore & Baker, 2001); because FrameNet is incomplete, a frame category was assigned to a group for any of its words that appear in FrameNet. When more than one frame was assigned to a group, all were retained. The groups associated with any of the communication frames and sub-frames were then extracted. In certain cases, the FrameNet "Communication" frame hierarchy had to be modified for our purposes: for example, lexical units described by the "Judgment-communication" frame are not distinguished for negative or positive valency (e.g., "acclaim" and "condemn" belong to the same FrameNet frame), which is obviously critical for historians exploring our data.[5] As a result, some manual adaptation of the FrameNet categories was required.

At present, we are representing communicative events using a simplistic scheme that assigns the role of *communicator* to the tagged PERSON or pronoun preceding the verb, and assumes the *topic* comprises the remainder of the sentence. This strategy works well for the Memoranda of Conversations, which typically exhibit a formulaic reporting structure in which the addressee is understood; for example, the text

> Mr. Kurusu asked whether this was our reply to their proposal for a *modus vivendi*. The Secretary replied that we had to treat the proposal as we did, as there was so much turmoil and confusion among the public both in the United States and in Japan.

yields the the following:

COMMUNICATOR: Mr. Kurusu
asked [ask : QUESTIONING: COMMUNICATION]
TOPIC: whether this was our reply to their proposal for a *modus vivendi*.
ADDRESSEE: Secretary Hull

COMMUNICATOR: The Secretary
replied [reply: COMMUNICATION_RESPONSE: COMMUNICATION]
TOPIC: that we had to treat the proposal as we did…
ADDRESSEE: Ambassador Kurusu

We are currently testing several freely-available parsers for English to provide a more reliable means to characterize communication events in the corpus. Because much of the language in the documents is stylistically complex, and in particular because of the cryptic syntax in the telegrams, we require a parser which is robust and (to some extent) forgiving. So far, the CMU Link Parser[6] appears to be the best choice for our data.

We are also exploiting role information in FrameNet to further refine entity references. For example, if "Japan" is the subject or object of a verb of communication, we can ascertain that this instance of "Japan" likely refers to the Japanese government (in the context of our documents, this would be the only possibility) and not the country of Japan; on the other hand, if "Japan" is the subject of a verb such as "attack" it is again likely to refer to the government, but as the object of "attack" it is more likely to refer to the country. This kind of distinction is critical for historical research: in the context of the Second World War, Japan "the government" and Japan "the country" are very different entities, and the ability of historians to distinguish the two is imperative.

The Ontology

We are currently building an ontology using RDF Schemas and OWL to describe our data. Where possible, we are extending existing general ontologies such as the OpenCYC/DAML[7] to include information relevant to the FDR data. For example, much of our information can be

[4] WordNet sense groupings were retained, although sense distinctions are not currently being considered.

[5] The FrameNet developers treat negative and positive valency as a semantic feature, but for our purposes individual frames are preferable.

[6] http://www.link.cs.cmu.edu/link/index.html

[7] http://www.cyc.com/

described by extending the upper ontologies for *government and military organizations* and *people related to organizations* provided in OpenCYC/DAML.

Creation of an ontology describing the entities in the FDR data demands considerable refinement, and in some cases re-definition of, categories provided in the OpenCYC/DAML ontologies. For example, our data includes references to a number of different types of *key-members* (defined in OpenCYC/DAML as someone who "is, or often gives input to, the organization's leader and thus may substantially influence the decisions of the organization"): these include not only government officials, but also members of the Japanese Imperial Family and various American personalities (e.g., Fred Kent, a New York banker, and E. Stanley Jones, a Methodist Minister), who are not government officials but either provide advice or, as in the case of certain members of the Imperial Family, act on the government's behalf in interactions with foreign ministers. Similarly, our data refer to entities such as Vichy France, Axis powers, ABCD powers, etc., whose classification according to existing OpenCYC categories is unclear; for example, France is a country, in Europe; but its status as a geo-political entity is complex: "Vichy France" is the official government of France in 1941, but governs only the southeastern portion of the physical territory commonly thought of as "France". It is not an Axis power, since it has only a "collaborative" relationship with the Germans, but neither is it an Allied power. The western and northern portion of France is considered "occupied France", governed by the Germans. Yet another governmental authority is what eventually comes to be called "Free France", Charles DeGaulle's counter-government located in London. To appropriately classify such entities, as well as the various official and unofficial documents, treaties, agreements, proposals, etc, it will be necessary to extend the types and properties of ontological categories provided in OpenCYC/DAML.

Our data provides a vast store of strategic, diplomatic, economic, and military information, and historians may approach the corpus with an interest in any one of them. Our ontology will therefore be necessarily tangled, since entities will participate in a variety of relations with one another. France again provides a good example: in terms of the country as a geo-political entity as a whole, it has obvious strategic importance due to its location within Nazi occupied Europe and proximity to Great Britain and Germany; its potential economic power; and its possession of several colonies in North Africa, Southeast Asia and the Caribbean. In 1941, these entities of "France", represented by the Vichy Government, were widely regarded as collaborationist. Within these entities, however, there were groups who chose to resist German domination and/or "France's" policy of collaboration. At the same time, the Free French based in London and led by DeGaulle, claimed that they were the true representatives of "France." At this stage in the war, however, only Vichy France held significant strategic importance, by virtue of its European territory, economic potential, limited Naval and other forces, and control of Indo-China and much of North Africa. In the light of this, the United States chose to maintain diplomatic relations with Vichy France, and refused to recognize any other organization that claimed to represent "France", while the British Government chose to recognize the Free French in London. Our ontology must represent all of these relationships, and at the same time allow for selective access, so that entities irrelevant to a particular view are not considered. All of this presents a fascinating challenge for ontology development, and, more generally, bears on the question of defining a "standard" ontology that can serve all interests and perspectives. Our experience so far suggests that this is neither possible nor desirable, and that mechanisms for selective ontology "views" need to be developed.

Conclusion

Because the FDR documents with which we are working deal with a narrow domain, they provide a unique opportunity to explore methodologies and representation issues to a level of detail not often addressed in previous work. The enhancement of the FDR/Pearl Harbor data will also provide scholars, educators and the general public with unprecedented access to a rich historical resource that may further advance our understanding of one of the most important events in American history.

The FDR/Pearl Harbor project is currently near its half-way point. In the next phase, we will be working intensively on building the ontology to support the data in the pilot corpus. In the final phase, the data will be made web-accessible and searchable, and documents will be linked to images of the originals. To implement intelligent search and retrieval based on inferencing over the ontologies, we will integrate one of the available inference engines, thus providing a "state of the art" resource for historical and political research.

Acknowledgements

This work is supported by U.S. National Science Foundation grant ITR-0218997. The authors would also like to thank the Franklin D. Roosevelt Presidential Library and Museum, and Stephen Rouch and David Meck for their assistance.

References

Cunningham. H. (2002) GATE, A General Architecture for Text Engineering. Computers and the Humanities, 36(2), 223—254.

Fillmore, C.J. & Baker, C. F. (2001). Frame Semantics for Text Understanding. Proceedings of WordNet and Other Lexical Resources Workshop, NAACL, Pittsburgh.

Ide, N., Bonhomme, P., & Romary, L. (2000). XCES: An XML-based Encoding Standard for Linguistic Corpora. Proceedings of the Second International Language Resources and Evaluation Conference. Paris: European Language Resources Association, 825-30.

Zhao, Y. & Karypis, G. (2001). Criterion functions for document clustering: Experiments and analysis. Technical Report TR #01--40, Department of Computer Science, University of Minnesota, Minneapolis, MN.