Who's Who in Your Digital Collection: Developing a Tool for Name Disambiguation and Identity Resolution

Carol Jean Godby, OCLC;
Patricia Hswe, University of Illinois at Urbana-Champaign (UIUC);
Larry Jackson, UIUC;
Judith Klavans, University of Maryland;
Lev Ratinov, UIUC;
Dan Roth, UIUC;
Hyoungtae Cho, University of Maryland

Abstract

In the past twenty years, the problem space of automatically recognizing, extracting, classifying, and disambiguating named entities (e.g., the names of people, places, and organizations) from digitized text has received considerable attention in research produced by the library, computer science, and the computational linguistics communities. However, linking the output of these advances with the library community continues to be a challenge. This paper describes work being done by the University of Illinois, the Online Computer Library Center (OCLC), and the University of Maryland to develop, evaluate and link Named Entity Recognition (NER) and Entity Resolution with tools used for search and access. Name identification and extraction tools, particularly when integrated with a resolution into an authority file (e.g., WorldCat Identities, Wikipedia, etc.), can enhance reliable subject access for a document collection, improving document discoverability by end-users.

Introduction

In the context of historical documents, the ability to find out who knew whom and why they were associated, in addition to whether the individuals are actually the ones the user is seeking, cultivates a potential for further, value-adding analysis of the documents' content. Discerning who's who in a digital resource collection is increasingly of interest to archivists, curators, and humanities scholars. The Perseus Digital Library has Named Entity Search Tools that mine its collections for people, places, and even dates.[1] The Metadata Offer New Knowledge (MONK) project offers a workbench for textual analysis on multiple levels, including a tool for recognizing and extracting named entities in its collections (which consist of works of eighteenth- and nineteenth-century American literature and works by William Shakespeare).[2] Named-entity extractors can also be found in cataloging utilities, such as the Computational Linguistics for Metadata Building (CLiMB) Toolkit[3], which addresses the "subject metadata gap" in visual resources cataloging by increasing subject access points for images of art objects.[4]

---

[1] See http://www.perseus.tufts.edu/hopper/.

[2] See http://www.monkproject.org/.

[3] http://www.umiacs.umd.edu/~climb/.

[4] Klavans, Abels, Lin, Passoneau, Sheffield, and Soergel 2009.

The problem of name disambiguation and identity resolution is made especially acute when many entities share the same name. Suppose a historian is seeking new insights about the assassination of John Kennedy. A Google search reveals that there are more than a few men named *John Kennedy;* the surname *Kennedy* itself is popular. The texts excerpted in Figure 1 describe about various *Kennedys*. To identify the relevant resources, the scholar would have to sift through search results one by one, a tedious task calling for automation. What would it take?

Document 1: "Composer and conductor John Kennedy is a dynamic and energetic figure in American music. Recognized for his artistic leadership, imaginative programming, audience development, and expertise in the music of our time, Kennedy has conducted celebrated performances of opera, ballet, standard orchestral and new music. His own compositions, from operas to chamber works, are praised for their new lyricism and luminous sound."[5]

Document 2: "In 1953, Massachusetts Sen. John F. Kennedy married Jacqueline Lee Bouvier in Newport, R.I. In 1960, Democratic presidential candidate John F. Kennedy confronted the issue of his Roman Catholic faith by telling a Protestant group in Houston, "I do not speak for my church on public matters, and the church does not speak for me.""[6]

Document 3: "John Kennedy was elected without opposition to his third term as State Treasurer in 2007. As Treasurer, he manages the state's $5 billion bank account including the investment of $3 billion in trust funds. He also oversees local and state bond issues and returns millions of dollars in unclaimed property each year. Prior to his position as Treasurer, Mr. Kennedy served as Secretary of the Department of Revenue, Special Counsel to Governor Roemer and Secretary of Governor Roemer's Cabinet."[7]

Fig. 1. Three texts about men named Kennedy.

The ideal software process would have to perform three tasks well enough to satisfy a discerning human judge. First, it would have to recognize the names. All name recognition software works by ingesting a string of text, such as the first sentence in the second document, and separating the names (Massachusetts, Sen. John F. Kennedy, Jacqueline Lee Bouvier, Newport, and *R.I.*) from the non-names (In, 1953, and married). This is a non-trivial task because the recognizer has to be smart enough to pick out names consisting of text strings that span more than one word, such as Jacqueline Lee Bouvier. It must also skip over the periods that indicate abbreviations (as in Sen. John. F. Kennedy or R.I.), but not those at the end of a sentence. Second, the recognizer must categorize the names. In the

[5] Quoted from the Web site about John Kennedy: http://www.johnkennedymusic.com/about.html. Retrieved December 15, 2009.

[6] Quoted from the Wikipedia entry on John F. Kennedy: http://en.wikipedia.org/wiki/John_F._Kennedy. Retrieved December 15, 2009.

[7] Quoted from the Web site for John Neely Kennedy: http://www.treasury.state.la.us/Home%20Pages/TreasurerKennedy.aspx. Retrieved December 15, 2009.

sample texts, all of the name strings containing the word Kennedy refer to people, although this will not always be true because the system will eventually encounter a text containing the organization name such as John F. Kennedy School of Government or a place name such as Kennedy Airport. It could also encounter strings that in some context are names, and in others are not, such as the first word in the sentence "Begin was the prime minister of Israel." Categorization effectiveness is a function of the diversity and extent of the training data supplied and of the algorithmic approach used. Finally, the software procedure must perform the most difficult task of all: assigning the real-world referents to the name strings. To help the scholar, the software would have to distinguish the John Kennedy from everyone and everything else named Kennedy, a task known as *name disambiguation* or *identity resolution*.

Because our project team has many librarians, we are interested in supporting research and scholarship like that of the hypothetical historian. An automated name recognizer paired with an identity resolver would support this goal and many others, including those that are central to the mission of libraries. For example, the output from these programs could be used to create more responsive interfaces for the discovery and retrieval of library materials. Or it could supply input to improved versions of resources that authoritatively describe the places, the people, and their inventions discussed in the published record, as well as the authors themselves.

Since there is no question that name recognition and identity resolution software would be key technologies for many applications enlisted in the service of preserving cultural memory, it is more interesting to ask why they haven't been pressed into service. The usual answer is that these programs, although incorporated to some degree in multiple commercial products, are not ready for full-scale deployment.[8] They may not be freely available or are difficult to use out of the box; processing time is too slow; the output has too many errors; and only name recognition, not entity resolution, is mature enough for serious consideration. But it's also undeniable that the output from these tools is already good enough for some library applications. To unleash their potential, researchers in the library community need to match this new technology with use cases that tolerate the current state of the art; form partnerships with the computer-science researchers engaged in front-line research in name recognition and identity resolution; and define realistic goals for future development.

To address these issues, we proposed the Extracting Metadata for Preservation (EMP) Project, funded by the National Digital Information Infrastructure and Preservation (NDIIPP) Program. As a collaboration among the University of Illinois at Urbana-Champaign, OCLC, and the University of Maryland, EMP researchers bring multidisciplinary perspectives from the library, computer science, and linguistics communities to the problem of high-quality identification and disambiguation of names. Our work has three goals: 1) to advance the state of the art in automated name identification and disambiguation; 2) to link the outputs of these programs to longstanding efforts in the library community to manage names and identities in the published record; and 3) to lower the barrier of access to these tools.

[8] Johnston 1990.

Related Research

Named Entity Recognition (NER) has been a key subject for researchers interested in accurate content extraction, information extraction, and information retrieval. Due to the centrality of personal names, places, dates, organizations and other named entities (NEs) in characterizing the topics in a document, audio or video clip, the quest for exactness in tokenizing these items has a long history. One of the earliest efforts to measure occurred at the Message Understanding Conferences (MUC), a series of workshops funded by the Defense Advanced Research Projects Association (DARPA). Projects funded by MUC participated in what are fondly called "computational linguistic bake-off's", where each system was run over a set of common data with results being submitted for evaluation by an independent set of evaluators through technology developed at the National Institute of Standards and Technology (NIST). The Named Entity task for MUC-6, held in 1995, consisted of three subtasks (entity names, temporal expressions, number expressions). The expressions to be annotated are "unique identifiers" of entities (organizations, persons, locations), times (dates, times), and quantities (monetary values, percentages). This task was intended to be of direct practical value (in annotating text so that it can be searched for names, places, dates, etc.) and an essential component of many language-processing tasks, such as information extraction.[9]

More recent approaches use a variety of techniques. In 2003 an overview of methods was provided at a workshop conducted by the annual Conference on Natural Language Learning (CoNLL)[10], supported by the Special Interest Group on Natural Language Learning of the Association for Computational Linguistics.[11] This reflects the current belief in the natural language processing and information extraction communities, that machine learning techniques, rather than programmed (rule-based) systems, are necessary in order to address the NER problem (and many other related problems).[12] Despite the emphasis on statistical machine learning techniques, most of the participants have attempted to use information other than the available training data, such as gazetteers and unannotated data.

The most frequently applied techniques in the CoNLL-2003 shared task were sequential classifiers of different sorts. At that time, one of the most poplar sequential classifiers was the Maximum Entropy Model (MEM), but several other sequential classifiers, such as Hidden Markov Models and Conditional Markov Models,[13] also were used. Many other machine learning approaches—including connectionist approaches, robust risk minimization, transformation-based learning, and support vector machines—were used for

[9] Grishman and Sundheim 1995.

[10] More information available at: http://www.cnts.ua.ac.be/conll2003/ner/. Also referred in this paper as the "CoNLL tagging scheme."

[11] Sang and De Meulder 2003.

[12] Klavans and Resnik 1992.

[13] Finkel, Grenager, and Manning 2005.

this problem, but it is clear today that architectural issues and features are the most important decisions, more than the specific training algorithm used.

One of the most complex tasks within the NER area is that of identifying nested entities. For example, Columbia University in the City of New York is an organization; however, the nested entity City of New York is a location, as is the entity nested within the nest, New York. Many corpus designers have chosen to avoid the issue of nesting entirelyand have annotated only the topmost entities. CoNLL,[14] MUC-6, and MUC-7 NER corpora, composed of American and British newswire, are all flatly annotated. A partial reason for this is that the NER task arose in the context of the MUC workshops, as small chunks of text which could be identified by finite state models or gazetteers. This then led to the widespread use of sequence models—first hidden Markov models, then conditional Markov models,[15] and, more recently, linear chain conditional random fields (CRFs).[16] None of these are able to model nested entities. Moreover, in essentially all sequential models it is often computationally difficult to represent non-local dependencies, which are often important in NER. This is one reason the approach used by Lev-Arie Ratinov and Dan Roth,[17] as described below, is not based on sequential classifiers but, rather, on state-of-the-art classifiers, which allows us to flexibly include non-local information.

The Name Extractor Tool

The EMP project uses the LBJ-based Named Entity Tagger[18], which was developed at the Cognitive Computation Group at UIUC by Ratinov and Roth.[19] The LBJ based NER was shown to be the best performing tool available today and its efficiency allows it to be used as part of applications that process large amounts of data. It extracts and labels non-nested named entities into four categories: locations (LOC), persons (PER), organizations (ORG), and miscellaneous names of human-created artifacts (MISC).

The algorithm incorporates a general model that learns from examples to identify named entities and classify them. It works in two stages. The baseline model makes a first cut by classifying the input text greedily left to right, using features that include, but are not limited to, the previous two tokens, the previous two classifications, and capitalization features. Most notably, the system does not use Part-Of-Speech tagging or shallow parsing information, which are common in other NER taggers. The second stage makes use of

[14] Sang and De Meulder 2003.

[15] Borthwick 1999.

[16] Lafferty et al. 2001.

[17] Ratinov and Roth 2009.

[18] Demo available at http://l2r.cs.uiuc.edu/~cogcomp/LbjNer.php.

[19] Ratinov and Roth 2009.

nonlocal features and features that exploit external knowledge. The classification model underlying the LBJ Named Entity Tagger is a regularized averaged perceptron algorithm.[20]

The two additional feature types added to the LBJ NER, along with other design decisions, account for its performance, which exceeds that of other state of the art tools and provides a necessary ability to adapt well to text from multiple domains and genres. Both feature types rely on automatically constructed evidence collected as part of the learning process. First, the system uses nonlocal features, such as the ratio of Named Entity types assigned to the current token previously in the text and context aggregation. By doing so, it makes use of the two-stage predication, where the first model is used to classify the text, while another model, similar in nature to the first, corrects the predictions to make them consistent within a document. Second, the system uses word class models and massive gazetteers automatically extracted from the online resource Wikipedia.

Consider, for example, the following text:

> SOCCER - [PER BLINKER] BAN LIFTED .
> [LOC LONDON] 1996-12-06 [MISC Dutch] forward [PER Reggie Blinker] had his indefinite suspension lifted by [ORG FIFA] on Friday and was set to make his [ORG Sheffield Wednesday] comeback against [ORG Liverpool] on Saturday. [PER Blinker] missed his club's last two games after [ORG FIFA] slapped a worldwide ban on him for appearing to sign contracts for both [ORG Wednesday] and [ORG Udinese] while he was playing for [ORG Feyenoord].

Fig. 2. Text displaying the annotated output of the LBJ Named Entity Tagger.

The system may incorrectly classify the first instance of Blinker at the first level of inference, but it will correct the prediction at the second level of inference by seeing that Blinker was a part of the expression Reggie Blinker, labeled as person. Furthermore, the system will use the knowledge extracted from Wikipedia, which states that Udinese, Sheffield Wednesday, Liverpool, and Feyenoord are football (soccer) clubs. The system will correctly label the second instance of Wednesday, since the expression Sheffield Wednesday was labeled as an ORG previously in the text. It is also important to note that the system uses the algorithm with large amounts of unlabeled text to abstract away words to a word class model, thus avoiding problems of data sparseness common in Natural Language Processing (NLP). For example, given the sentence, "FIFA slapped," the system knows that slapped is used in similar contexts as "devised, reimposed, manifested, commissioned, authorised, imposed, etc," helping the system to label *FIFA* as ORG.[21]

---

[20] A perceptron is an "On-line, mistake driven, additive update rule. Perceptron updates the weights in a target node by adding to them a learning rate that is a function of the type of mistake made (either positive or negative) and the strengths of features in the example" (Carlson et. al. UIUC technical report; http://l2r.cs.uiuc.edu/~cogcomp/software/snow-userguide/node43.html).

[21] More details may be found in Ratinov and Roth 2009.

In addition to the extensive evaluation described in the CoNLL 2009 presentation by Ratinov and Roth, we also assessed how well the LBJ Named Entity Tagger performs in comparison with other state-of-the-art name extractor applications used in the library community. Besides the LBJ tagger employed in our project, two other tools were assessed: ClearForest Gnosis (ClearForest)[22], which is a FireFox add-on application that semantically processes Web pages, linking named entities to further information about them; and the Stanford Named Entity Recognizer (NER), developed by the Stanford Natural Language Processing Group using a Character-based Maximum Entropy Markov Mode (MEMM), which is implemented in Java.[23] For the additional evaluation, we selected five text samples taken from diverse domains, ran the samples through each tool, and compared the raw performance of the results. We also engaged a human annotator to tag named entities in each text sample and compared the human-generated results with those obtained from evaluation of the aforementioned three NER tools. It is important to note that in all cases addressed here the tool was evaluated on text taken from domains that are vastly different from the domain it was trained on. In principle, when one wants to use such a tool in a different domain, the best course of action is to re-train the tool on the target domain. The results here, therefore, should also be taken as evidence of the robustness and adaptability of the tool.

For mentions that were exactly matched, the F-scores for the LBJ tagger on the five text samples ranged from 47.83% to 78.99%, depending on the domain; for partially matched mentions, the F-scores ranged from 60.13% to 85.71%. The closest competitor, ClearForest, had F-scores for exactly matched mentions that ranged from 36.14% to 61.73%; for partially matched mentions, the F-scores for ClearForest ranged from 42.77% to 75.86%. In the version evaluated, the LBJ NER tool was tuned to yield the best F1 score, which is the harmonic average of recall and precision, although it is possible to tune it to emphasize one over the other. In general, a high precision rate is often important in dealing with extremely large collections, since the latter would be likely to yield more errors, and thereby waste the user's time. High recall rates, though, reflect the coverage of the tool—the percentage of entities identified—and are desirable where the search must be exhaustive, such as in research or legal applications. In general, with the version evaluated ClearForest had slightly higher precision but significantly lower recall (that is, it identified significantly fewer entities). One lesson from this evaluation that we intend to act on is to simplify the ability of a user to retrain the LBJ NER tool on a target domain, and to allow a user to easily trade recall and precision.

Resolving Identities

As we said in the Introduction, entity recognition is only the first part of the problem of capitalizing on the rich information associated with names in unstructured text. The second is identity resolution: determining which person, place, or concept in the real world the

---

[22] Available at: https://addons.mozilla.org/en-US/firefox/addon/3999.

[23] Available at: http://nlp.stanford.edu/software/CRF-NER.shtml. The Stanford NER is also part of the evaluation reported by Ratinov and Roth in their CoNLL 2009 paper.

extracted name refers to. This is a classic problem in the philosophy of language.[24] In a nutshell, identity resolution requires the help of an authority who can step outside the text and link the name with the appropriate referent—such as a mother who names her child *John Fitzgerald Kennedy*, a public official who witnesses this act, or a journalist who writes about it. This link then needs to be fixed so that it remains constant over time, persisting even into eras when the named entity has passed out of living memory. Thus, if the name-referent link is robust, 23rd-century readers of a book published in 1966 about the assassination of John Fitzgerald Kennedy will understand that the book is about the American president who was elected in 1960, just as their counterparts in the 20th century did.

Since the creation of a name-referent link is a vexing problem for philosophers and is occasionally challenging for human readers, it would appear intractable for a software algorithm that does not have access to the world beyond a set of input texts. Except for the people, places, and things encountered in their everyday experience, humans don't have this access, either. But they still manage to understand texts like those excerpted in Figure 1. We can infer that since relatively few people are personally acquainted with the composer, the 35th American president, or the state treasurer of Louisiana (the examples presented in Figure 1 above), they grasp the meaning of these texts by consulting identity resolution authorities—textbooks or other works of nonfiction, documentary films, encyclopedias, or their own memories of these works—who describe the identity behind the name in enough detail to establish a proxy reference.

Algorithms that attempt to resolve identities also consult a resolution authority to establish the identities of the various people named Kennedy in texts such as the ones we have described. Stated more formally, the problem to be solved has three parts. First, name occurrences are extracted from the text, such as John Kennedy, or simply Kennedy. Second, a software process must match the name occurrences against those found in an identity resolution authority. This task is easy if the name occurrence is unusual and has only one entry in the authority. But more typically, the name is ambiguous and has multiple representations, which makes a third step necessary: generating candidates from the identity resolution authority and selecting the correct one, a task that usually requires that the input text be mined for clues about the identity of the name occurrence, such as birth and death dates for personal names, or city and country names for places.

So what is a good identity resolution authority for a software process? Computer scientists argue that Wikipedia is appealing because it is a high-quality edited text that is freely available. It has a relatively large coverage (over two million entities as of August 2009) and is frequently updated by human annotators who enhance the hyperlink structure. In particular, the most important named entities mentioned in Wikipedia articles are linked to the corresponding Wikipedia pages, which are also annotated with a list of human-created categories. These features allow us to obtain statistics, such as how often a given set of tokens refers to a given Wikipedia page; how often two Wikipedia concepts appear in the same Wikipedia page; and how the texts are associated with abstract Wikipedia categories. These statistics permit the construction of expressive disambiguation models. Ratinov and Roth are developing a disambiguation system that assigns the correct Wikipedia entries to named entities and concepts identified in blogs and texts retrieved by standard information

[24] Kripke 2000.

retrieval algorithms.[25] Their system builds on the work of researchers who attempted to enrich the hypertext structure of Wikipedia by expanding the list of named entities that link to the corresponding articles.[26]

The librarians on the EMP team have proposed the use of library authority files for identity resolution. Typically created by national libraries to establish unambiguous references to the people, places, and topics represented in the published record, library authority files are highly encoded and designed for machine processing. Figure 3 shows a portion of the record for John Fitzgerald Kennedy from the Library of Congress Name Authority File. The various fields in the record supply birth and death dates, alternative forms of his name, associated subjects, and the coded names of the agencies that vouch for the accuracy of this information.

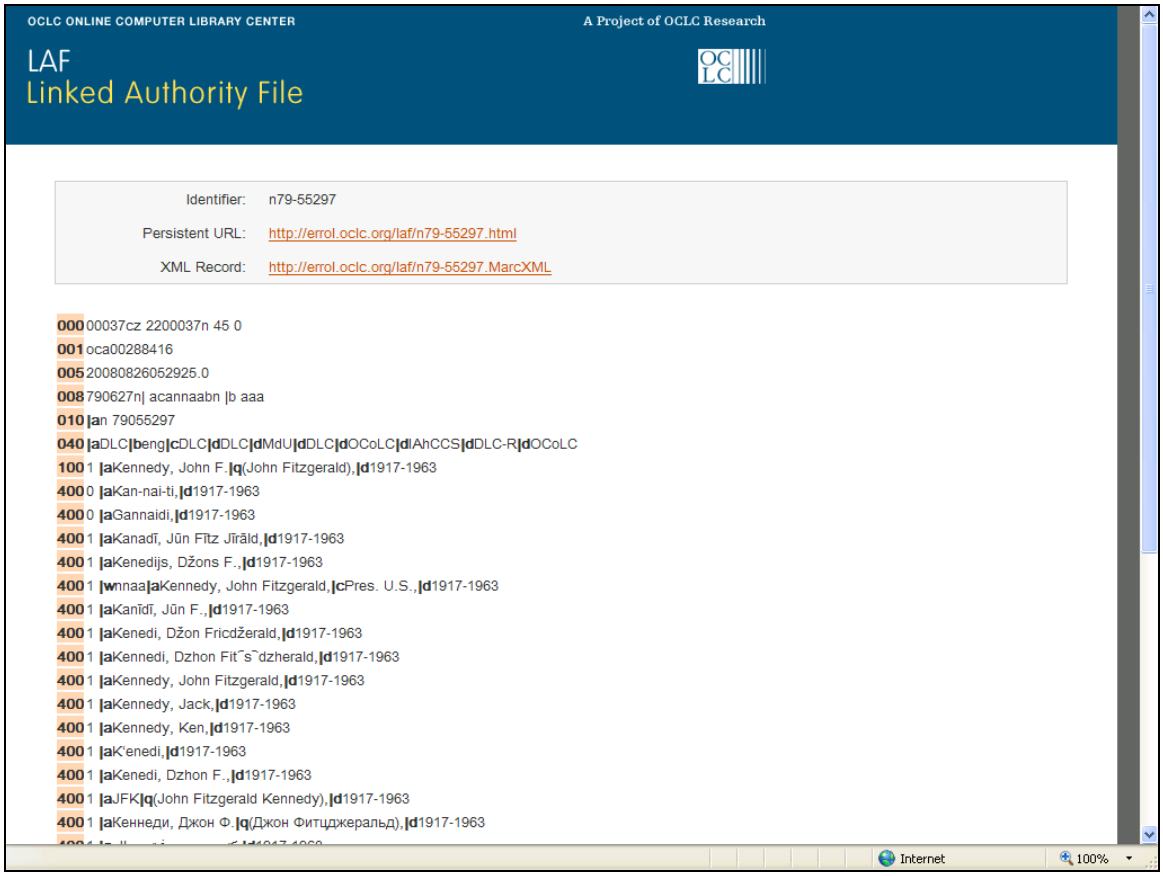

Fig. 3. The Library of Congress Authority record[27] for John Fitzgerald Kennedy

[25] UIUC Tech Report, Dec. 2009.

[26] See, for example, Cucerzan 2007, or Mihalcea and Csomai 2007.

[27] Available at: http://errol.oclc.org/laf/n79-55297.html.

In the past five years, classification experts in the library community have recognized the need to create authority files that span national and linguistic boundaries. One outcome is the Virtual International Authority File[28], a collaborative effort that merges authority files from thirteen national libraries. Another example is OCLC's WorldCat Identities[29], a Web-accessible collection with 27 million pages about personal names,[30] which have been populated with links and other data obtained from multiple authority files, Wikipedia, and collections of bibliographic records—in particular, OCLC's database of 158 million records representing records contributed by 71,000 libraries worldwide. Since these resources are automatically compiled, they must also rely on identity resolution algorithms that extract name occurrences and select the correct identity from a list of candidates. But since the authority file data is highly encoded and the scope is restricted to names represented in the published record, it is relatively easy to discover distinctive information such as the names of works an author has published. In the next section, we discuss an extended example that illustrates the use of authority files for identity resolution.

At present, the EMP project team is debating how to reconcile these two approaches to identity resolution. The team's computer scientists argue that the library authority files contain data that is too sparse for algorithms tuned for the rich unstructured text of Wikipedia. Or that Wikipedia is comprehensive, while the library authority files are restricted to the published record. It is also clear, however, that the two types of resources are complementary. If the goal is to identify the names of authors extracted from text obtained from the open Web, the correct resolution is more likely to come from WorldCat Identities than from Wikipedia, which currently has fewer than 125,000 articles about authors. At the same time, WorldCat Identities can be probably be enhanced by algorithms that work on unstructured text: they promise to locate authors who are well-known and influential yet not represented in the published record, since they speak only through blogs or Web sites that have gone viral.

Library Applications of Named Entity and Identity Resolution Software

OCLC's interest in the technology developed in the EMP project stems from the need to link unstructured text to its large collections of highly coded records, such as bibliographic and authority records, and other metadata required to support the management and discovery of library resources. OCLC researchers are now turning their attention to the many streams of full text that are associated with these materials, such as author biographies, reader reviews, online reference works, unpublished or pre-published manuscripts collected in institutional repositories, and similar materials. In the terminology developed in the problem statement above, the association of unstructured text to structured metadata is necessary, because the coded material often has the identities, while the unstructured text is what mentions one form of the name.

---

[28] Available at: http://viaf.org/

[29] Available at: http://orlabs.oclc.org/Identities/

[30] WorldCat Identities also has 7 million pages about corporate names and 14,000 subject names. (Ralph LeVan, personal communication)

Consider an example from QuestionPoint, the virtual reference service maintained by OCLC in partnership with the Library of Congress.[31] Library patrons submit a question through the QuestionPoint interface, which is automatically routed to the closest participating librarian, based on the IP address of the computer from which the question originates. The librarian answers the question in a response window after a time delay that varies from a few minutes to a few days. Questions and answers that are of general interest are eventually collected in a database, which users can search and browse. Figure 4 shows one example of a question-answer record, which is a full-text document. If readers want to find out more about the broad topic, John Fitzgerald Kennedy, or the authors of the books cited in the librarian's answer, they may associate this record to other resources at OCLC and elsewhere. But they would have to cut and paste selected text into WorldCat.org, Google, Wikipedia, or other resources that might provide more depth or context. The interface doesn't do this work for them. In other words, this text frequently mentions names, but identity resolution is up to the reader.

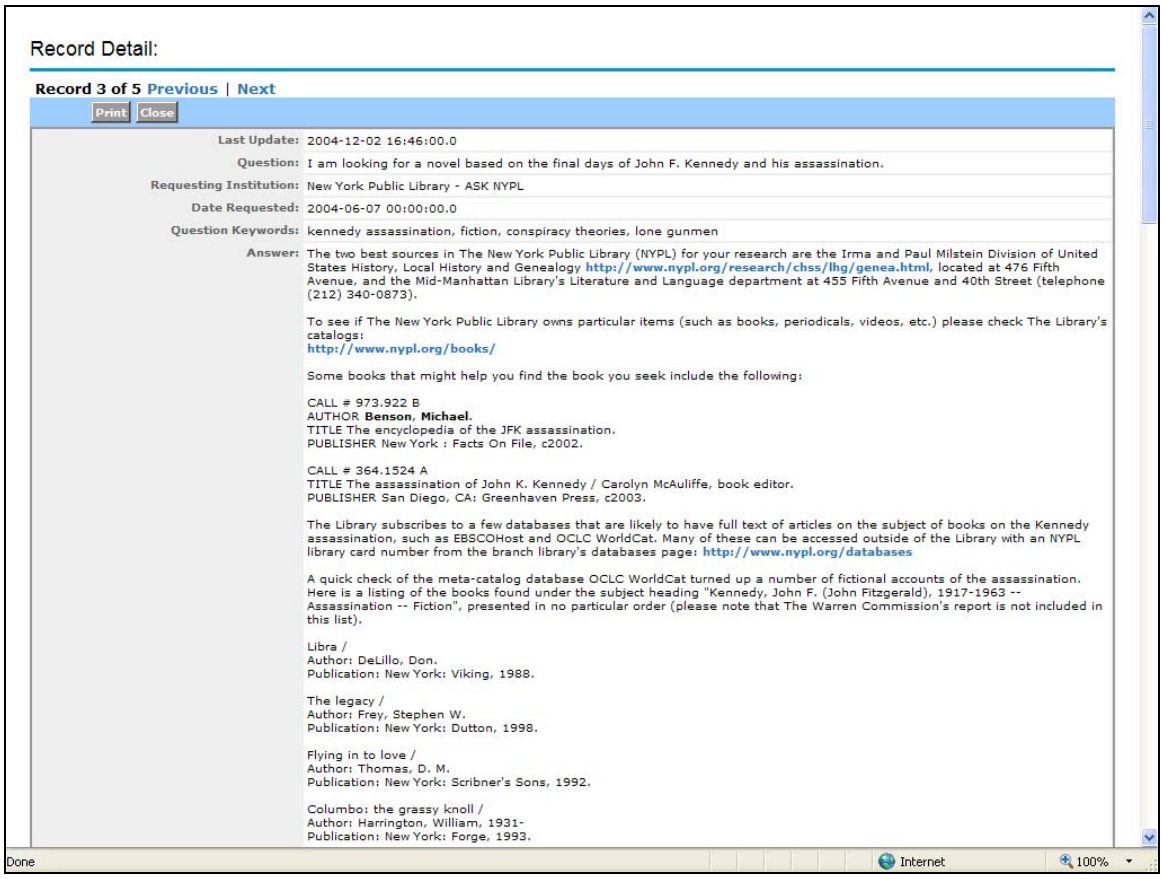

Fig. 4. A record in the QuestionPoint knowledge base.

With more sophisticated information extraction from unstructured text and algorithms that link the output to structured resources, the records in this database could be enhanced to

[31] http://www.oclc.org/services/brochures/211401usb_questionpoint.pdf

add clickable links to the QuestionPoint record. When these links become available, the reader would, with minimal effort, be able to find *The Encyclopedia of the JFK Assassination* or *The Assassination of John F. Kennedy* in his/her local library, find a list of other books by the author Michael Benson or the editor Carolyn McAuliffe (listed in Figure 4 above), and discover other works about the assassination of John Fitzgerald Kennedy, or related broader and narrower topics. Since the structured metadata already supports such exploration, the only missing piece is the association with texts such as the QuestionPoint answer. The EMP tools are designed to provide this information.

The first step is to run the QuestionPoint record through the LBJ Named Entity Tagger to obtain the name occurrences. The results are shown in Figure 5. Organizational names are green, locations are blue, personal names are bright red, and miscellaneous names are brownish red.

To see if The [ORG New York Public Library] owns particular items (such as books, periodicals, videos, etc.), please check the library's catalog.

Some books help you find the book you seek include the following:

CALL # 973.922 B
AUTHOR [PER Benson, Michael].
TITLE The encyclopedia of the JFK assassination.
PUBLISHER [LOC New York]: [ORG Facts On File], c2002.

Fig. 5. NER markup for a fragment of a QuestionPoint answer.

In initial tests with the QuestionPoint answer records, the most important problem is the parsing and linking of the book citations, shown here. To obtain useful output from the NER tool, we had to overcome some built-in bias and train it to recognize names of the form [PER Last, First] and [PER Last, First Initial]. With about 450K of training data, we obtained results that recognized these new forms while retaining the tool's native ability to recognize names conforming to the more usual [PER First Last] pattern. The training data also specifies that any name following the pattern [LOC] and a colon (:) is an organization, leading to the correct recognition of publisher names. The title remains untagged, but it is recognized through a regular-expression match as the text that intervenes between the pattern [PER Last, First] and [LOC]:[ORG], as shown.

Once the name occurrences have been extracted and selected, the next step is to link them to the correct identities. The obvious tool for accomplishing this goal is the Wikipedia tool being developed by Ratinov and Roth, which enables linking the name occurrences to Wikipedia, but this turns out not to be useful. Although Wikipedia has an entry for Michael Benson, the name annotated in Figure 5 above, it describes the documentary filmmaker, not the author of *The Encyclopedia of the JFK Assassination,* the title annotated above in Figure 5. The deeper problem is that Wikipedia is not the best identity resolution authority for the

task of assigning clickable links to book citations, because it contains relatively few articles about authors.

WorldCat Identities is a more promising authority. The page for Michael Benson, the author of *The Encyclopedia of the JFK Assassination*, is shown in Figure 6. This page has a rich collection of links for this author, including a list of his published books, alternative forms of the author's name, a list of co-authors (with indirect links to their published works), and a list of subject headings associated with the author.

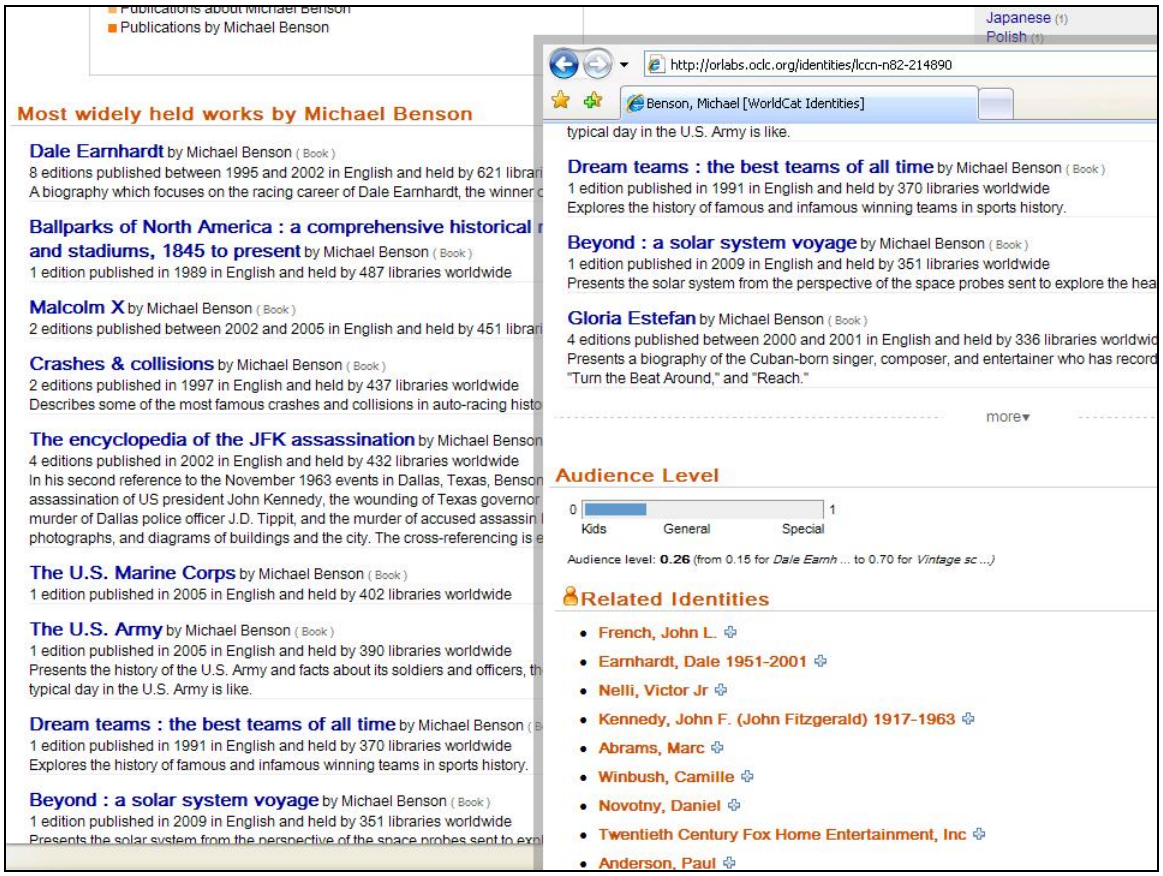

Fig. 6. The WorldCat Identities page for Michael Benson.

WorldCat Identities is created algorithmically, primarily by collecting data from OCLC's WorldCat database. Preprocessing utilities mine WorldCat's bibliographic records, creating a separate page for every author, as well as for every person (real or fictitious) who has been the subject of a published work. But in a database the size of WorldCat, there are many authors named Michael Benson. How does the algorithm link to the correct author?

The answer turns out to be elegantly simple. The key insight is that the name of the author and the title of the book can be thought of as a bigram, in which the first element is Michael Benson and the second is Encyclopedia of the JFK Assassination. Significantly, an author-title bigram is highly improbable and often unique. In other words, it is unlikely that more

than one Michael Benson authored a book with this title about the JFK assassination. Since WorldCat Identities can be searched from an API that accepts an Open URL, a publicly accessible specification for representing information typically found in a bibliographic record,[32] the author and title can be sent in the form shown here:

http://worldcat.org/identities/find?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:identity&rft.namelast=Benson&rft.namefirst=Michael&rft.title=MICHAEL+BENSON+AND+THE+ENCYCLOPEDIA+OF+THE+JFK+ASSASSINATION+%28+%27.

This URL triggers a fuzzy-name search against WorldCat Identities, which returns a results list containing a list of 49 Michael Bensons. The top-ranked Benson, the correct link, goes to the Identities pages shown above in Figure 6. To finish the task of presenting clickable answers to QuestionPoint queries, a software routine embeds this intelligence into the XML of the text that is served through the user interface.

This example shows that in a best-case scenario, the problem of associating book citations found in full text with a link that disambiguates the author's name can reduce to the problem of name recognition. Once the name occurrences have been correctly extracted from the input text, sophisticated search and ranking algorithms already in place generate the candidate identities and recommend the correct one.

Other problems at OCLC involving links between resources resemble the QuestionPoint example, but it is instructive to make the underlying issues more explicit. In the example we have discussed, the name occurrence is in the unstructured text and the identity is in a collection of structured resources, which constitute an identity resolution authority. There may be more than one identity resolution authority, which may have complementary strengths. The task of disambiguating the name of a book author is best accomplished by referring to an identity resolution authority that is customized for the published record. However, if the task is to establish the identities of names of local historical or cultural figures, about whom little or nothing has been published, Wikipedia may be a better authority than WorldCat Identities. These observations imply that identity resolution algorithms will perform better when multiple resources can be consulted. It is a priority for future work to determine how this is best accomplished.

Yet a more significant issue emerges from this data. What happens when *no* available name resolution authority can resolve a name occurrence? A name would still be extracted from unstructured text, along with other identifying characteristics, such as a book title, if the name is an author; birth and death dates, if the person is famous; a subject domain associated with the person's work, and so on. But if no match can be made even against a detailed text, the text itself now contains one form of name occurrence as well as important clues for resolving the identity. If these clues are collected, they could form a valuable first draft for a larger and more timely identity resolution resource that is populated automatically, a huge improvement over the current state of the art.

Conclusion

[32] Van de Sompel and Beit-Marie 2001.

The QuestionPoint exercise is a simple proof-of-concept demonstration for a set of processes that start with the automatic extraction of names from unstructured input text and end with significant enhancements to a commercially available product. This is a work in progress, however. The most immediate need is for improved recognition of the large variety of book and article citation styles in text that was not designed for machine processing. Similar problems are being addressed by other researchers at OCLC who are using the NER tool to extract names from text fields in a bibliographic record, with the goal of increasing the navigable links in collections of published works.[33]

In fact, there is no shortage of uses for robust NER extraction and identity resolution utilities in the library community. A name extractor tool can also be used to parse names that occur in collections of digitized government documents. But it will have to be expanded to recognize not only the names of persons, locations, and organizations, but also government information applications, position titles, edifices, geographic features, geo-political regions, and laws or regulations. Once found, these names can provide searchers many more precise access points into collections than are currently available through state-of-the-art systems.

Key persons, places, concepts, and artifacts occur in information retrieval in almost all disciplines, making progress on the identity resolution problem a broad, cross-cutting need. Outside library circles, identity resolution authorities would need to be created from scratch. Large numbers of topically focused communities have literatures emerging on the Web, thanks in no small part to prototypes and best practices developed under IMLS and Library of Congress research funding.

In the next phase of development we will address the disambiguation of recognized names resulting from such software. We plan to run our named entity extraction software on a variety of directory-like Web pages[34] as a means of facilitating the initial construction of name authority files, with an eye to establishing "community-authored" authority lists. The University of Illinois has done extensive work in archiving digitized state government documents, resulting in a vast collection of materials. Notoriously rich in name variations, these digital government materials would support this stage of investigation extremely well. Efficient citizen (and government staffer) access into that corpus would benefit considerably from name disambiguation.

References

[33] See, for example, the demo of Work Records In WorldCat, accessible at: http://frbr.oclc.org/research/pages/026336461.html. The field named 'Derived Terms' was populated with the LBJ-NER tool.

[34] Official webpage lists and rosters like the State of Illinois Telephone Directory (http://www.illinois.gov/teledirectory/printable.cfm) published by the Governor's office, or the Illinois General Assembly's list of Illinois State Senators (http://www.ilga.gov/senate/) are available, although socially contributed Web lists such as Wikipedia pages may similarly be processed.

Anderson, D. 1991. Automatically generated references in minimal-level authority records. *Information Technology and Libraries* 10(4): 251-262.

Borthwick, A. 1999. A maximum entropy approach to named entity recognition. Thesis. New York University. 1999.

Brown, P.F., P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18(4): 467–479.

Carlson, A., C. Cumby, J. Rosen, and D. Roth. The SNoW learning architecture. Technical report, 1999.

Chinchor, N. 1997. MUC-7 named entity task definition. *In Message Understanding Conference Proceedings MUC-7.* http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ne_task.html.

Cucerzan, S 2007. Large-scale named entity disambiguation based on Wikipedia. In EMNLP 2007: *The Joint meeting of the Conference on Empirical Methods in Natural Language Processing*

Davis, P. T., D. K. Elson, and J. L. Klavans. 2003. Methods for precise named entity matching in digital collections. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries. Houston, Texas, May 27 - 31, 2003,*125-127. Washington, DC: IEEE Computer Society.

Epstein, S. B. 1985-1986. Automated authority control: a hidden timebomb? Parts I and II. *Library Journal* 110(18): 36-37; and 111(1): 55-56.

Finkel, J.R. and C. D. Manning. 2009. Joint parsing and named entity recognition. In *Proceedings of the North American Association for Computational Linguistics – Human Language Technology (NAACL HLT) 2009,* 326-334. Boulder, CO.

Finkel, J.R., T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs Sampling. *43rd Annual Meeting of the Association for Computational Linguistics. Proceedings of the Conference,* 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf.

Grishman, R. and B. Sundheim. 1995. Design of the MUC-6 evaluation. In *Proceedings of the 6th Conference on Message Understanding. Columbia, Maryland, November 06 - 08, 1995,* 1-11. Morristown, NJ: Association for Computational Linguistics.

Hickey, T. 2008. VIAF and WorldCat Identities. European Library Automation Group (32nd ELAG Library Systems Seminar). Wageningen, The Netherlands. http://library.wur.nl/WebQuery/file/formulier/profielelaglt_i00117278_001.ppt.

Jernigan, E. R. T. 1939. Authority files and official catalogs. *Library Journal* 64: 23-25.

Johnston, S. H. 1990. Desperately seeking authority control: automated systems are not providing it. *Library Journal* 115(16): 43-46.

Klavans, J.L., E. Abels, J. Lin, R. Passonneau, C. Sheffield and D. Soergel. 2009. Mining texts for image terms: the CliMB project. In *Digital Humanities 2009. Conference Abstracts. University of Maryland, College Park, USA. June 22 – 25, 2009,* 184-186. College Park, MD: Maryland Institute for Technology in the Humanities (MITH). http://www.umiacs.umd.edu/~jimmylin/publications/Klavans_etal_DH2009.pdf.

Klavans, J. L. and P. Resnik, eds. 1996. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* Cambridge, MA: MIT Press.

Knight, K. and S. Luk. 1994. Building a large-scale knowledge base for machine translation. In *Proceedings of the 12th National Conference on Artificial Intelligence, Seattle, WA, USA, July 31 - August 4, 1994.* Menlo Park, CA: AAAI Press; Cambridge, MA: MIT Press.

Kripke, S. 2000. *Naming and Necessity.* Oxford: Blackwell.

Lafferty, J., A. McCallum and Pereira, F. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Eds., C. E. Brodley and A. Pohoreckyj Danyluk. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*. San Francisco: Morgan Kaufmann.

Li, X. and Roth, D. 2005. Discriminative training of clustering functions: theory and experiments with entity identification. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL). 29-30 June 2005. University of Michigan, Ann Arbor, Michigan, USA*. Madison, WI: Omnipress, Inc.

Mihalcea, R. and A. Csomai. 2007. Wikify! Linking documents to encyclopedic knowledge. In CIKM 2007: ACM Sixteenth Conference on Information and Knowledge Management.

Ratinov, L. and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL-2009: Thirteenth Conference on Computational Natural Language Learning*. 147-155. Boulder, CO: Association for Computational Linguistics. http://aclweb.org/anthology-new/W/W09/W09-1119.pdf.

Renear, A. H., K. M. Wickett, R.J. Urban, D. Dubin, S. L. Shreeves. (2008). Collection/item metadata relationships. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*. http://dcpapers.dublincore.org/ojs/pubs/article/viewPDFInterstitial/921/917.

Roth, D. and W. Yih. 2007. Global inference for entity and relation identification via a linear programming formulation. In *Introduction to Statistical Relational Learning,* eds. L. Getoor and B. Taskar, 553-580. Cambridge, MA: MIT Press.

Sang, E. F. T. K. and De Meulder, F. 2003. Introduction to the CoNLL-2003 shared task: language independent named entity recognition. In *Proceedings of Seventh Conference on Natural Language Learning (CoNLL). 31 May-1 June 2003. Edmonton, Canada,* 142-147. Association for Computing Machinery: Morristown, NJ.

Taylor, A. G. (1984). Authority files in online catalogs: an investigation of their value. *Cataloging & Classification Quarterly* 4 (3): 1-17.

Tillett, B. B. 1989. Considerations for authority control in the online environment. *Cataloging & Classification Quarterly* 9 (3): 1-11.

Van de Sompel, H. and O. Beit-Marie. 2001. Open linking in the scholarly information environment using the OpenURL framework. *D-Lib Magazine,* 7 (3). http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html.

Wilson, D. R. 2005. Name standardization for genealogical record linkage. Family History Technology Workshop (FHTW 2005). Brigham Young University. http://fht.byu.edu/prev_workshops/workshop05/FHTCD/session3/s3-randall_wilson_NameStandardization.pdf.