Name This! Automating Metadata Extraction through a Named Entity Recognition Tool

Jean Godby - OCLC
Judith Klavans - University of Maryland
Dan Roth - University of Illinois at Urbana-Champaign
Patricia Hswe - University of Illinois at Urbana-Champaign
Presentation Outline

• The problem of named entity recognition

• Collaboration among three institutions
  – University of Illinois at Urbana-Champaign
  – University of Maryland
  – OCLC
Three problems

For each name in a stream of text:

• **Recognize it**
  – Distinguish names from non-names
  – Assign it to a broadly recognized category
    • Personal, corporate, geographic

• **Cluster it**
  – Associate variants of the same name

• **Assign an identity**
  – … or the name’s real-world referent
  – Select the canonical form of a name
Document 1: The Justice Department has officially ended its inquiry into the assassinations of John F. Kennedy and Martin Luther King Jr., finding "no persuasive evidence" to support conspiracy theories, according to department documents. The House Assassinations Committee concluded in 1978 that Kennedy was "probably" assassinated as the result of a conspiracy involving a second gunman, a finding that broke from the Warren Commission's belief that Lee Harvey Oswald acted alone in Dallas on Nov. 22, 1963.

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." JFK was known for his liberal approaches to religion and civil rights.

• **Recognize it**
  – *John F. Kennedy*
  – [PERS John F. Kennedy]
  – [PERS David Kennedy]

• **Cluster it**
  – *John F. Kennedy, Kennedy*
  – *David Kennedy, Kennedy*
  – John F. Kennedy != David Kennedy

• **Assign an identity**
  – John F. Kennedy, President and Senator
  – David Kennedy, Poet and Author
EMP: Extracting Metadata for Preservation

- NDIIPP National Digital Information Infrastructure and Preservation Program
- Funded by the Library of Congress
- Grant to University of Illinois at Urbana-Champaign.
- University of Maryland and OCLC are subcontractors.
EMP project goals

- Lower the barrier of access to high-end Named-Entity Recognition tools
- Create tools for open use
- Demonstrate use in digital library applications
Project deliverables

1. Use named entity recognition (NER) toolkit developed at the University of Illinois
2. Address the three problems with named entities.
   – automatically extract proper names from text
   – normalize them
   – develop algorithms to assign identities
3. Evaluate with human subjects.
4. Create a Web service interface for the NER tool.
5. Demonstrate the NER toolkit in library applications.
6. Deliver the toolkit as an open source software package.
Three Institution Collaboration

1. **Named Entity Recognition (NER)**
   - University of Illinois at Urbana-Champaign

2. **Computational Linguistics for Metadata Building (CLiMB)**
   - University of Maryland

3. **Research projects on names management**
   - OCLC
The Named Entity Recognizer

• Developed by Dan Roth at UIUC (Cognitive Computation Group, Department of Computer Science)

• Implements a machine learning solution to named entity recognition

• Demo available at: http://l2r.cs.uiuc.edu/~cogcomp/LbjNer.php
The input sentences are:

- George Washington (February 22, 1732[1][2][3] – December 14, 1799) was the first President of the United States, (1789-1797)[4] and led the Continental Army to victory over the Kingdom of Great Britain in the American Revolutionary War (1775–1783). Washington was chosen to be the commander-in-chief of the American revolutionary forces in 1775. The following year, he forced the British out of Boston, lost New York City, and crossed the Delaware River in New Jersey and defeated the surprised enemy units later that year. As a result of his strategy, Revolutionary forces captured the two main British combat armies at Saratoga and Yorktown. Negotiating with Congress, the colonial states, and French allies, he held together a tenuous army and a fragile nation amidst the threats of disintegration and failure. Following the end of the war in 1783, Washington retired to his plantation on Mount Vernon, prompting even King George III to name him "the greatest man in the world."[5][need quote]

The tagged output is:

- [PER George Washington] (February 22, 1732 (1) (2) (3) – December 14, 1799) was the first President of the [LOC United States] (1789 – 1797) and led the [ORG Continental Army] to victory over the [ORG Kingdom] of Great [LOC Britain] in the [MISC American Revolutionary War] (1775 – 1783). [LOC Washington] was chosen to be the commander-in-chief of the [MISC American] revolutionary forces in 1775. The following year, he forced the [MISC British] out of [LOC Boston], lost [LOC New York City], and crossed the [LOC Delaware River] in [LOC New Jersey] and defeated the surprised enemy units later that year. As a result of his strategy, Revolutionary forces captured the two main [MISC British] combat armies at [LOC Saratoga] and [ORG Yorktown]. Negotiating with [ORG Congress], the colonial states, and [MISC French] allies, he held together a tenuous army and a fragile nation amidst the threats of disintegration and failure. Following the end of the war in 1783, [LOC Washington] retired to his plantation on [PER Mount Vernon], prompting even King [PER George III] to name him "the greatest man in the world." (5) [need quote]
Under development for EMP

- New research on identity resolution

- Increase usability
  - Web service interface
  - Workflows for training
  - Incorporating data into demos

- Performance evaluation and tuning
  - Comparisons with human taggers
  - Comparisons with other algorithms
  - Evaluation of scoring methodologies
Three Institution Collaboration

1. **Named Entity Recognition (NER)**
   - University of Illinois at Urbana-Champaign

2. **Computational Linguistics for Metadata Building (CLiMB)**
   - University of Maryland

3. **Research projects on names management**
   - OCLC
Clip of the John F. Kennedy Funeral
Looking for Bobby Kennedy
CLiMB
Computational Linguistics for Metadata Building

• Use text associated with images
• Assist cataloger in identifying high quality subject descriptors
• Uses a range of computational analysis tools
• Named Entities are linked to existing resources
The Problem:

Insufficient Subject Access to Images
Current Process Flow

Image Catalogers

Catalog Record

Record from AICT website

Minimal metadata for image, no descriptive terms.

Image Searchers
The famous painted limestone bust of Akhenaton’s queen, Nefertiti (fig. 3-33), exhibits a similar expression of entranced musing and an almost mannered sensitivity and delicacy of curving contour. The piece was found in the workshop of the queen’s official sculptor, Thutmose, and is a deliberately unfinished model very likely by the master’s own hand. The left eye socket still lacks the inlaid eyeball, making the portrait a kind of before-and-after demonstration piece. With this elegant bust, Thutmose may have been alluding to a heavy flower on its slender stalk by exaggerating the weight of the crowned head and the length of the almost serpentine neck…
The famous painted limestone bust of Akhenaton’s queen, Nefertiti (fig. 3-33), exhibits a similar expression of entranced musing and an almost mannered sensitivity and delicacy of curving contour. The piece was found in the workshop of the queen’s official sculptor, Thutmose, and is a deliberately unfinished model very likely by the master’s own hand. The left eye socket still lacks the inlaid eyeball, making the portrait a kind of before-and-after demonstration piece. With this elegant bust, Thutmose may have been alluding to a heavy flower on its slender stalk by exaggerating the weight of the crowned head and the length of the almost serpentine neck...
**Catalog Record with Enhanced Subject Metadata**

**Main Entry: Category or Artist**

Egypt (New Kingdom)

**Title or Description**

Queen Nefertiti: (front view), ca. 1350 B.C. [Dynasty XVIII]

**[Location] - AICT Photo ID #**

Aegyptisches Museum, Berlin
#EN010

**CLI MB Terms**

- Limestone [carbonate rock], aat: 300011286
- Basts [figures], aat: 300047457
- Akhenaton
- Queens [monarchs], aat: 300025483

[more CLiMB terms]

**Text References**

- Adams 3: pl. 5.22 [alt.]
- Adams 4: pl. 5.22 [alt.]
- Adams AAT: pl. 4.42 [alt.]
- Adams AAT2: pl. 3.38 [alt.]
- Gardner 11: pl. 3-33
- Gardner 12: pl. 3-33
CLiMB Approach

Techniques exceed simple keyword indexing by:

1. applying advanced semantic categorization to text segments
2. identifying coherent phrases
3. associating terms with a thesaurus, and
4. applying disambiguation algorithms to these terms.
The CLiMB Interface

As Europeans came to understand the world through printed maps and geographies, landscape emerged as a popular subject. The Italian artist and biographer Giorgio Vasari noted that "there is no cobler's house without its landscape because one becomes attracted by their pleasant view and the working of depth." Here the torture of Saint Catherine is overwhelmed by the scenery in a way typical of the panoramic "world landscapes" painted by northern artists. Perspective and point of view are manipulated to provide the most information possible: this is a God's eye view that sees everything simultaneously. Varied terrain and captivating detail compel the viewer to travel across the picture with his eyes. This painting may be the work of Matthias Cock, whose brother Hieronymus was a well-known publisher of prints, including many by Pieter Bruegel. Notice the distinctive mountain crags here and similar ones in other landscapes nearby painted by...
Under development for EMP

• Compare performance of UIUC name extractor tool and CLiMB’s name extractor

• Evaluation and usability studies

• Research on scoring evaluation
Three Institution Collaboration

1. Named Entity Recognition (NER)
   • University of Illinois – Urbana Champagne

2. Computational Linguistics for Metadata Building (CLiMB)
   University of Maryland

3. Research projects on names management
   • OCLC
OCLC

• Our needs
  – Subject access to the published record
  – Improved authority files
  – Links to unstructured and un-curated text

• Our resources
  – WorldCat Identities
  – Linked authority files
  – Growing collections of unstructured text
Matches for 'kennedy, john f.'

1. Kennedy, John F. published: 1940-2008 (Children of presidents--United States)
2. Kennedy, John F. (John Fitzgerald) 1917-1963 (Presidents--United States)
8. Kennedy, John F. 1942-
Matches for 'kennedy, david'

1. Kennedy, David published: 1854-2008
2. Kennedy, David M. published: 1939-2008
6. Kennedy, David 1954-
8. Kennedy, David M. 1958-
9. Kennedy, David 1959-
10. Kennedy, David W. 1948-
11. Kennedy, David 1943-
12. Kennedy, David 1969-

Results Key
- A Personal Identity
- A Personal Identity from a controlled vocabulary
- A Corporate Identity
- A Corporate Identity from a controlled vocabulary

Text Size indicates relative popularity of an Identity

(SUBJECT) The information in parentheses indicate the Subject Heading associated with that Identity
David Kennedy (disambiguation)

From Wikipedia, the free encyclopedia

David Kennedy may refer to:

- David Kennedy (1955-1984), American writer and son of politician Robert Kennedy
- David Kennedy (Australian politician) (born 1940), politician and member of the Australian House of Representatives
- David Kennedy (guitarist) (born 1976), American rock musician
- David Kennedy (jurist) (born 1954), American legal academic and scholar of international law, vice president of international affairs at Brown University, formerly of Harvard
- David Kennedy (actor), British actor
- David Kennedy (advertising) (born 1939), American advertising executive and co-founder of Wieden+Kennedy advertising agency
- David Kennedy (film producer), producer of Saving Milly and the upcoming Dark Shadows film adaptation
- David Kennedy (New Zealand) (1864-1936), New Zealand priest, astronomer and educator.
- David M. Kennedy (1915-1996), American businessman, economist and former United States Secretary of the Treasury
- David M. Kennedy (historian), Pulitzer Prize-winning historian and professor at Stanford
- David T. Kennedy (born 1934), U.S. politician
- Dave Kennedy (born 1953), Irish racing driver

This disambiguation page lists articles associated with the same personal name. If an internal link led you here, you may wish to change the link to point directly to the intended article.

Categories: Human name disambiguation pages
Research questions

• How do we establish reliable links between structured and unstructured resources?

• How do we expedite the automated processing of WorldCat Identities data?

• How do we populate new authority records?
SPONSORSHIP With the consent of the affected members, [ORG Representative Froehlich] was removed as principal sponsor, and [ORG Representative Feigenholtz] became the new principal sponsor of HOUSE BILL 4995. With the consent of the affected members, [ORG Representative Soto] was removed as principal sponsor, and [ORG Representative Ryg] became the new principal sponsor of SENATE BILL 222. HOUSE RESOLUTION The following resolution was offered and placed in the [ORG Committee] on Rules. HOUSE RESOLUTION 1009 Offered by Representative [PER Madigan] [LOC WHEREAS], [PER

FutureGen - Wikipedia, the free encyclopedia - Windows Internet Explorer

FutureGen
From Wikipedia, the free encyclopedia
(Redirected from FutureGen)

This article or section contains information about a proposed, planned, or existing construction power station.
It may contain information of a speculative nature and the content may change as the construction approaches completion of the power station approaches, and more information becomes available.

FutureGen is a US government project announced by President George W. Bush in 2003[1]; its initial plan involved the construction of a near zero-emissions coal-fired power plant to produce hydrogen and electricity while using carbon capture and storage.[2]
..and an opportunity

• Use WorldCat Identities as a resource for establishing the referent of a name

• Build on research at UIUC that uses Wikipedia for identity resolution

• Utilize domain specific proper name resources (e.g. Union List of Artist Names ULAN from Getty)
Matching names to identities

Extract Names from Text

David Kennedy
Kennedy

Search WorldCat Identities

Kennedy, David 1954-
Kennedy, David M. 1958-
Kennedy, David 1959-
Kennedy, David W. 1948-
Kennedy, David 1943-

Identifying information

born...1959—Leicester, England
The New Poetry

A unique identifier

Match identity info from text
And Identities pages
John F. Kennedy

Extract Names from Text

John F. Kennedy
Kennedy, JFK

Search WorldCat Identities

Identifying information

Married to Jacqueline Lee Beauvier

A unique identifier

Match identity info from text
And Identities pages

Kennedy, John F. published: 1940-2008
(Children of presidents--United States)
Kennedy, John F. (John Fitzgerald)
1917-1963 (Presidents--United States)
Kennedy, John F. (John Fitzgerald)
(Children of presidents--United States)
John F. Kennedy Center for the Performing Arts (U.S.) published: 1959-2008
Project status

- Evaluating annotated data
- Developing workflows
- Demo building
- Integrating NER tool with CLiMB
- Proof-of-concept results of identity resolution algorithms
- Building partners
References

• NDIIPP site: http://ndiipp.uiuc.edu/
• CLiMB (Computational Linguistics for Metadata Building) site: http://www.umiacs.umd.edu/~climb/
• Named Entity Recognizer (NER) UIUC homepage: http://l2r.cs.uiuc.edu/
• OCLC’s WorldCat Identities: http://www.worldcat.org/identities/
EMP Project Team - 1

• Jean Godby, Project Lead, OCLC
  • godby@oclc.org
• Judith Klavans, Project Lead, UMD
  • jklavans@umd.edu
• Dan Roth, Project Lead, UIUC
  • danr@illinois.edu
• Patricia Hswe, NDIIPP Project Manager, UIUC – phswe@illinois.edu
EMP Project Team - 2

- Sarah Dotson, User Studies, UMD
  - sjdotson@gmail.com
- Larry Jackson, Collection Consultant, UIUC – lsjackso@illinois.edu
- Rebecca LaPlante, Project Coordinator, UMD – laplante@umd.edu
- Lev-Arie Ratinov, Graduate Research Assistant, UIUC - ratinov2@uiuc.edu
- Devon Smith, Software Engineer, OCLC
  - smithde@oclc.org
Summary
Thank you.

Questions?
Notation and Task Definition

• 'Mention' (def) a string in text that refers to a real-world entity; 'Entity' (def) is the entity being referred to.

• The goal: given unlabeled text, link mentions to entities. ...

Document 1: The Justice Department has officially ended its inquiry into the assassinations of John F. Kennedy and Martin Luther King Jr., finding "no persuasive evidence" to support conspiracy theories, according to department documents. The House Assassinations Committee concluded in 1978 that Kennedy was "probably" assassinated as the result of a conspiracy involving a second gunman, a finding that broke from the Warren Commission's belief that Lee Harvey Oswald acted alone in Dallas on Nov. 22, 1963.