Looking Ahead to Person Resolution

2004 Family History Technology Workshop
March 25, 2004

Mary D. Taffet (mdtaffet@syr.edu)
Syracuse University
School of Information Studies
Center for Natural Language Processing

Presentation Outline

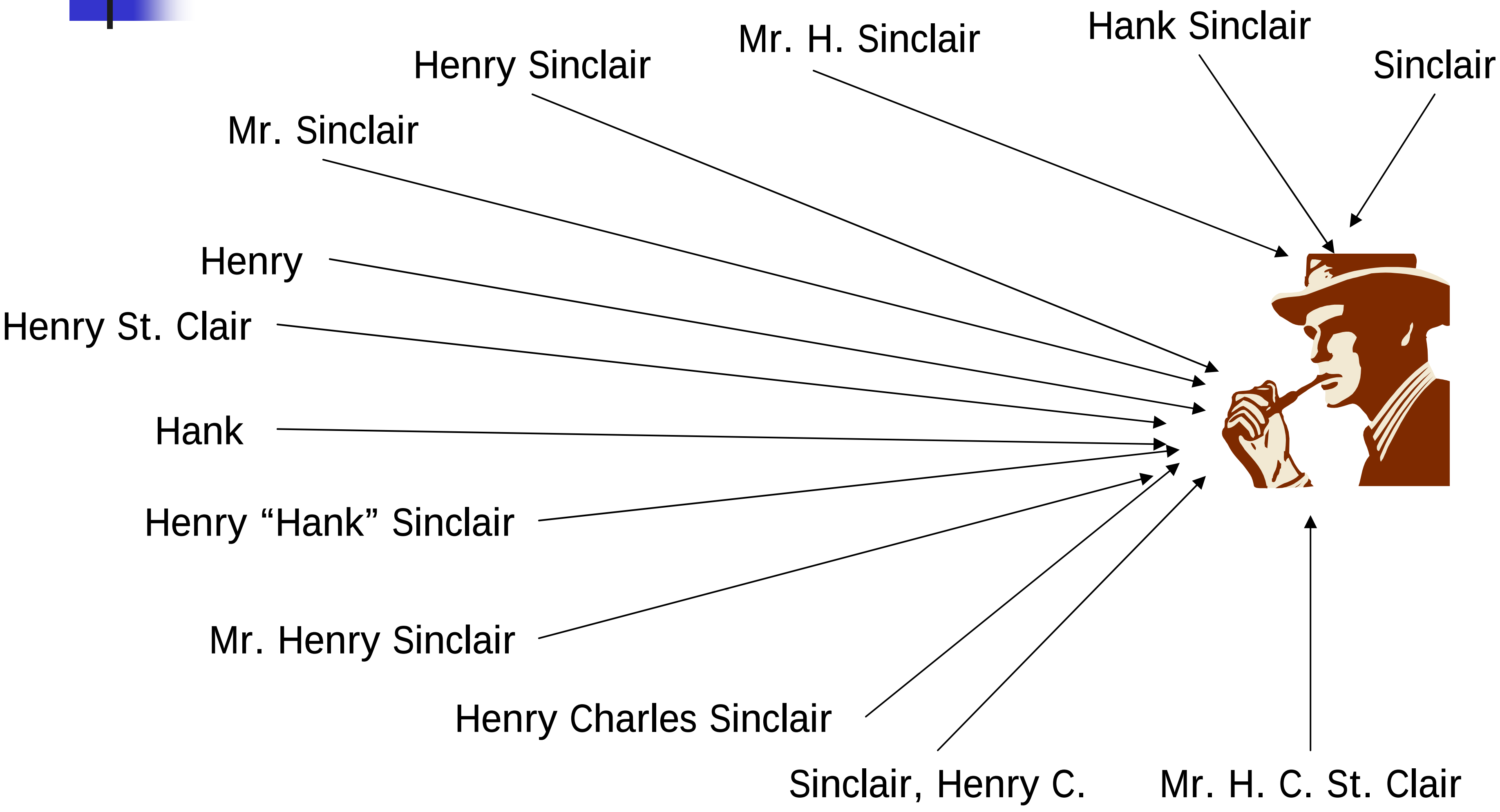
- Background
- Goal of the study
- Research Design
- Methodology
  - Phase I – User Study I
  - Phase II – Design & Implement Person Resolution Algorithm
  - Phase III – Evaluation & User Study II

Background

- Document understanding and retrieval with regard to names of people is hard because person names are very prone to ambiguity.

- Most difficult form of ambiguity is inherent in the many-to-many relationship between person names and people
  - Many-to-many relationship can be decomposed into two separate relationships.

Many names-one person
Multimorphic Person Names



Hank Sinclair

Mr. H. Sinclair

Henry Sinclair

Sinclair

Mr. Sinclair

Henry

Henry St. Clair

Hank

Henry “Hank” Sinclair

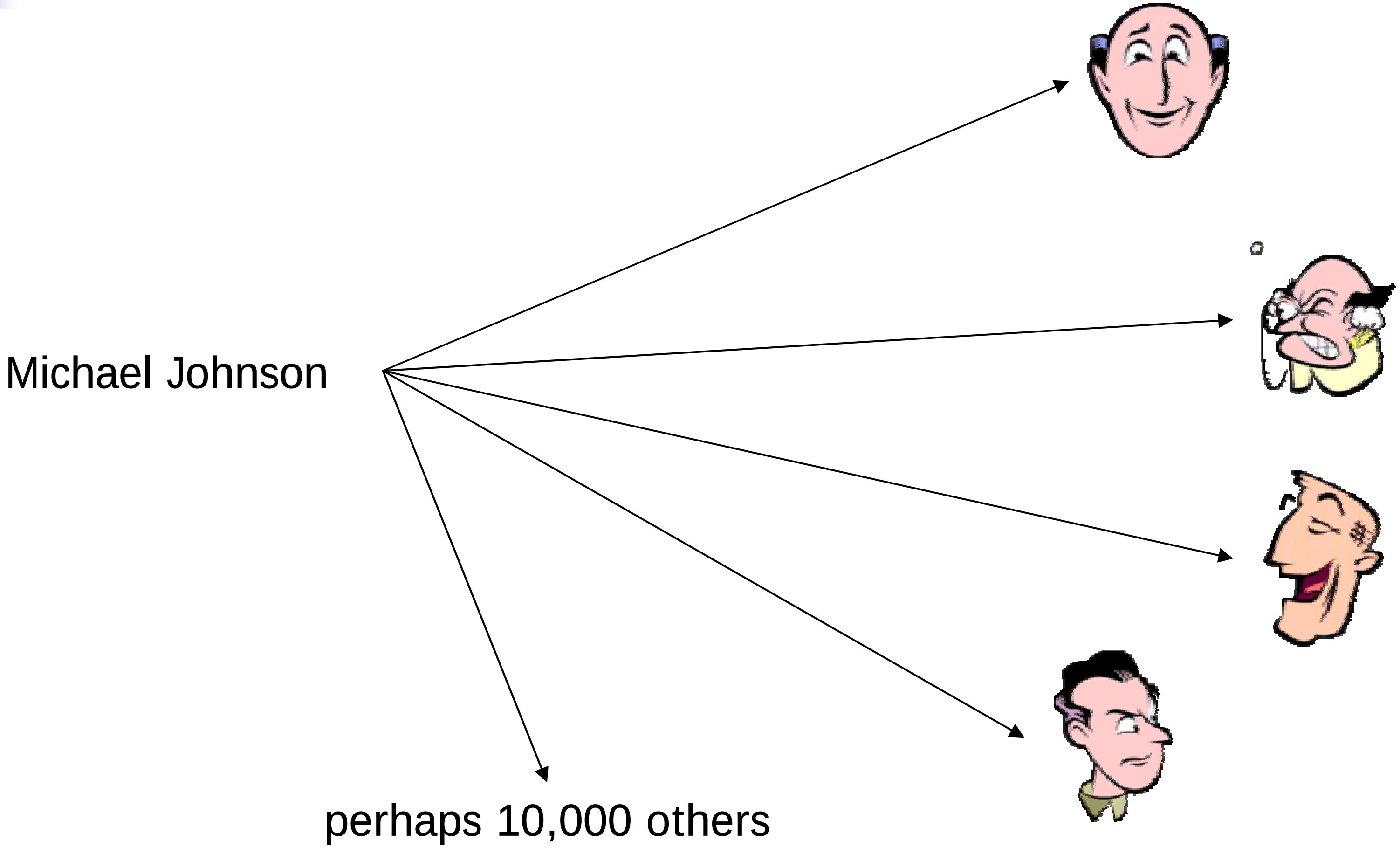
Mr. Henry Sinclair

Henry Charles Sinclair

Sinclair, Henry C.

Mr. H. C. St. Clair

One name-many people Multireferent Person Names



Michael Johnson

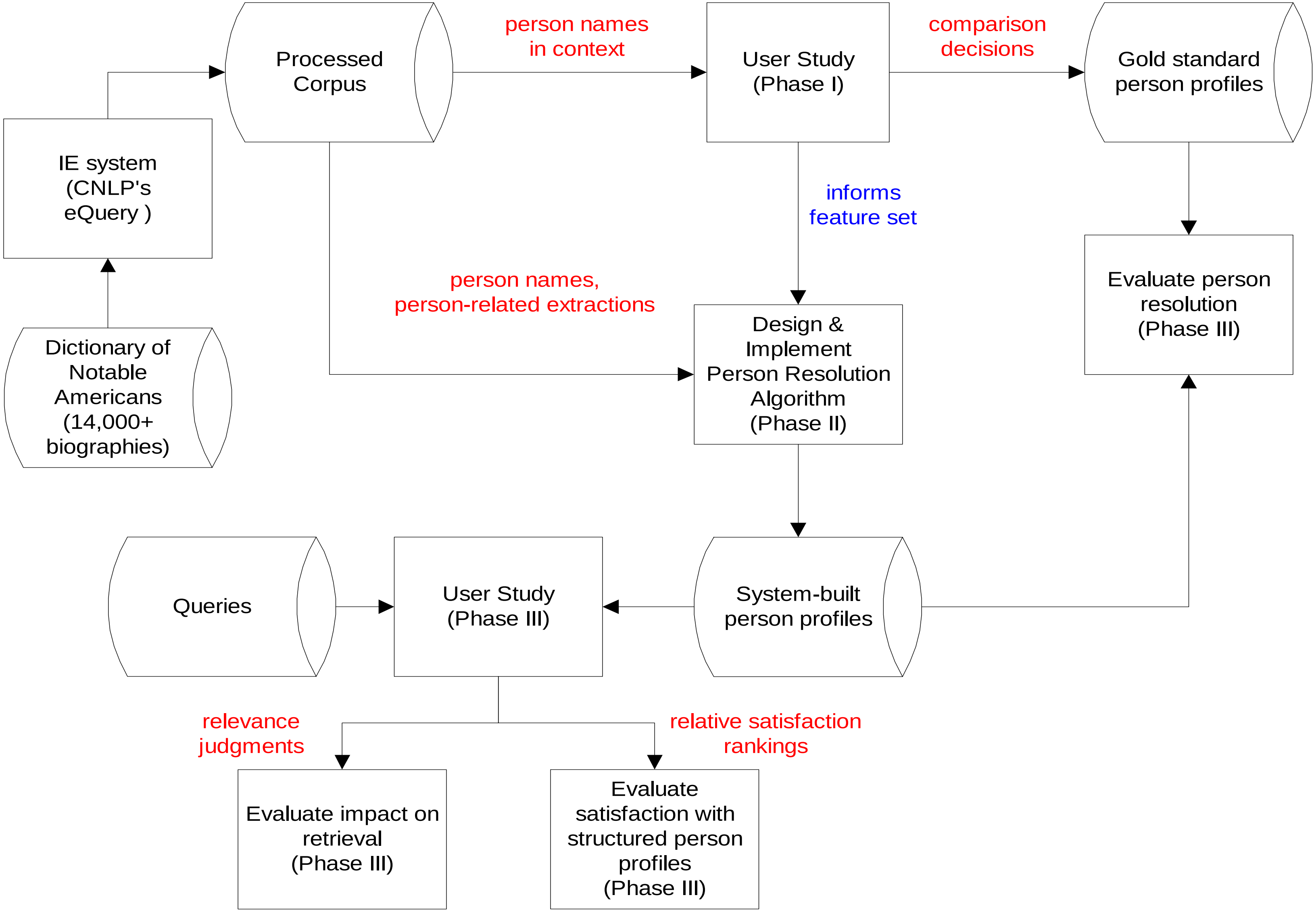
perhaps 10,000 others

Goal of the Study

- Seek a solution to the person name understanding and retrieval issues due to the existence of multireferent person names and multimorphic person names.

- Method: Creation of person profiles through a process called person resolution
  - Assignment of multireferent person names which refer to different people to different person profiles
  - Assignment of multimorphic person names (i.e. name variants) which refer to the same person to the same person profile

Research Design


Processed Corpus
person names in context
User Study (Phase I)
comparison decisions
Gold standard person profiles
IE system (CNLP's eQuery)
informs feature set
Evaluate person resolution (Phase III)
person names, person-related extractions
Dictionary of Notable Americans (14,000+ biographies)
Design & Implement Person Resolution Algorithm (Phase II)
Queries
User Study (Phase III)
System-built person profiles
relevance judgments
relative satisfaction rankings
Evaluate impact on retrieval (Phase III)
Evaluate satisfaction with structured person profiles (Phase III)

Methodology

- Corpus-based study
  - Corpus = Dictionary of Notable Americans (1904) supplied in electronic form by Ancestry.com; 14,000+ biographical narratives
  - Document processing using eQuery system developed by the Center for Natural Language Processing (CNLP)
    - Named entity recognition (bracketing & categorization)
    - Coreference resolution for singular personal pronouns and person-related noun phrases (definite and indefinite)
    - Extraction of person-related information useful for person resolution
  - Creation of person profiles via person resolution
- Three phases:
  - Phase I – User study I
  - Phase II – Design & Implement Person Resolution Algorithm
  - Phase III – Evaluation & User Study II

Phase I – User Study I Part A

- Human Judgments Captured Online
    - Genealogists will be shown two person names in context
    - Genealogists will decide if the two names refer to the same person or not
    - Web-based survey created and pretested; needs some adjustment
        - Adjustment will consist of having genealogists read both documents in their entirety before showing them which names they are to judge
    - Once adjustment made and tested, web-based survey will be made available online and advertised to genealogists at large
- Goals
    - Short term:
        - Creation of gold standard person profiles by researcher based on these decisions
    - Long term:
        - Creation of reusable test collection

Phase I – User Study I Part B

- Human Judgments Captured In-person
  - Teach-back method
    - Knowledge elicitation technique involving knowledge engineer and domain expert with goal of capturing expert’s knowledge
    - Will modify this method to have researcher be an observer only; expert genealogist will interact with a novice genealogist
  - Session flow
    - Experts and novices will be given document pairs to read (as in Part A)
    - Expert will decide if the two names in context refer to the same person or not (as in Part A)
    - Expert will explain the basis for their decision to the novice
    - Novice will teach back to the expert until the expert is satisfied that the novice understands
- Goal
  - Capture textual and real-world information which might be useful as part of the feature set for person resolution

Phase II – Design & Implement Person Resolution Algorithm

- Cyclic/iterative process trying out different combinations of input features, processes, and preliminary outputs
  - Input
    - Person-name/extraction pairs to be resolved
    - Features to use for person resolution
  - Process
    - Clustering
    - Record Linkage, probably Probabilistic
      - Adapted to unstructured documents
    - Decision Tree
    - possibly Support Vector Machines
  - Preliminary Output
    - Classification labels
    - Probabilities for match/non-match
    - Groupings or clusters
- Final Output is Person profiles

Phase III – Evaluation & User Study II

- Evaluation of Person Resolution Algorithm
  - Intrinsic evaluation
    - Comparison to gold standard person profiles (not the same ones used during design phase)
    - B-Cubed metric developed by Amit Bagga for named entity resolution
  - Extrinsic (task-based) evaluation
    - IR experiment based on retrieval of documents
    - User Study II, Part A
      - Gather human judgments about relevance of documents to queries; web-based survey

Phase III – Evaluation & User Study II

- Evaluation of Person Profiles
  - User Study II, Part B
    - Gather human judgments about relative satisfaction with results in the form of:
      - Undifferentiated list of ranked documents
      - List of documents minimally formatted to show results of person resolution (header with person name, birth date, death date, etc.)
      - Structured person profiles with links to associated documents