Looking Ahead to Person Resolution

Mary D. Taffet[1]

School of Information Studies
Syracuse University
Syracuse, NY 13244-4100
{mdtaffet@syr.edu}

Introduction

There is a many-to-many relationship between person names and people. This many-to-many relationship can be decomposed into two separate relationships. There is a many-to-one relationship between person names and people in that a person may be referred to by a multitude of names that vary from each other in terms of structure, order, and/or spelling. For example, a person who is usually referred to as *John Smith* might also be referred to as *Mr. Smith, Mr. John Smith, Mr. J. Smith, John, Smith, Jack Smith, John H. Smith*, or *John Henry Smith*, seemingly ad infinitum; this same person might also be referred to as *John Smiht* if somebody made an error. I call this type of ambiguity multimorphic (many forms/shapes) ambiguity. There is also a one-to-many relationship between person names and people. There are at least two persons named *George Bush*, and there were at least two persons named *John Kennedy*. There could be hundreds of thousands of persons named *John Smith*. I call this type of ambiguity multireferent ambiguity.

This ambiguity of person names presents a problem when processing documents automatically for the purposes of information retrieval or information extraction. Ideally, the variant forms of one person's name should be brought together, and the multiple referents for a single name should be teased apart. Solutions to this problem are important for several domains, one of which is genealogical research.

Person Resolution

The primary goal of the study described here is creation of a person resolution algorithm which will automatically resolve both multireferent person name ambiguity and multimorphic person name ambiguity within and across full-text documents. Person profiles will be created as a byproduct of this process; person profiles will contain at a minimum all name instances that have been resolved to the same person, and will probably contain other elements such as important events.

The corpus chosen for this study is Biographies of Notable Americans [1], which is based upon a prior print publication [2]. Ancestry.com supplied an electronic copy of the corpus for this study. The corpus has been installed, and 14,520 biographical narratives have been extracted from it. Preliminary processing will locate, mark, and uniquely identify all instances of person names throughout the corpus. Subsequent processing will resolve pronouns, definite noun phrases, and indefinite noun phrases that refer to person names within documents, such that

[1] The author is a Research Analyst at the Center for Natural Language Processing, which is part of the School of Information Studies. A dissertation proposal based on the study described here was defended January 14, 2004.

each such expression will be assigned the id of a unique person name instance. Where person names within documents are incomplete, such as name instances consisting of only a first name or only a last name, an effort will be made to recover the full form of the name where possible. Generic extraction will be performed on the documents to extract not only the person names, but also other entities, events, and the relationships between them. The entities, events and relations will be stored in a semantic case frame representation, which will be used to identify person-related extractions. Person-related extractions are those entities, events and relations that appear in the same case frame as a person name. The person-related extractions provide the context that will be used to resolve multireferent and multimorphic person name ambiguity.

The proposed study will be conducted in three major phases. Phase one involves a user study to elicit human decisions about whether two names refer to the same person, as well as the basis for those decisions. A web-based instrument was created and pretested for this user study, but adjustments are needed before the web-based instrument can be made available online. Human subjects were shown two documents side-by-side, and one person name in each document was highlighted. The subjects were asked to compare the two names and then choose one of the following responses:

- The two names **DEFINITELY** do **NOT** refer to the same person; instead they refer to two different persons
- The two names **PROBABLY** do **NOT** refer to the same person
- The two names **POSSIBLY** do **NOT** refer to the same person
- There is **not enough information available** to decide whether the two names refer to the same person
- The two names **POSSIBLY DO** refer to the same person
- The two names **PROBABLY DO** refer to the same person
- The two names **DEFINITELY DO** refer to the same person
- There is an **error in the data**; one or both of the items highlighted above is not the name of a person

In making this decision, the pretest subjects often made false assumptions based on incorrect pronoun resolution. For instance, in the following sentence, the name *General Buell* was highlighted:

> “He joined **General Buell**'s army in February, 1862, and for his conduct at Shiloh, where he was wounded, he was made brigadier-general of volunteers, receiving his commission after he had gained greater honors at the siege of Corinth and at Perryville, where he commanded a brigade.”

Pretest subjects reading this sentence sometimes assumed incorrectly that it was *General Buell* who was wounded and *General Buell* who was made brigadier-general, but instead it was the subject of the biography (*Edward Henry Hobson*) to whom these two actions were truly ascribed. The pretest subjects had incorrectly mentally resolved the pronoun “he” to *General Buell* in these two instances. Once a satisfactory solution to this pronoun resolution problem is found, the web-based instrument will be released online and advertised to genealogists. In addition to collecting information about the basis for the decisions using the web-based instrument, one or more in-person group sessions will be held. During these group sessions, subjects will use teaching to share their knowledge about how to decide if two names refer to the same person. It is anticipated that very rich data can be collected during these group teaching sessions. The results from phase one will include (a) the comparison decisions, which will be used to create gold standard person profiles, and (b) the basis for those decisions, which will be used in part to derive the feature set used for person resolution.

Phase two includes design, testing, and implementation of a person resolution algorithm. An algorithm is like a system in that it involves some form of input, some form of processing, and some form of output. Multiple design cycles will be involved, with each design cycle encompassing a different set of input features, a different approach, and possibly a different set of outputs. The inputs will include person names and person-related extractions, along with a set of input features. The input features will most likely fall into two major groups; one group will likely include attribute-value pairs, such as *gender=male*, and the other group will likely include the results of small comparisons, such as the result of a comparison between birth dates for example. The outputs might include classification labels, probability estimates, or perhaps clusters. If classification labels are used, the values would likely be “yes” or “no”, with “yes” indicating that two names refer to the same person (i.e. should be assigned to the same person profile) and “no” indicating that the two names do not refer to the same person (i.e. should be assigned to different person profiles). If probability estimates are used, then there would likely be two estimates – a match probability, along with a non-match probability. A high degree of match coupled with a low degree of non-match would be interpreted as meaning that two names refer to the same person (i.e. should be assigned to the same person profile); conversely, a high degree of non-match and a low degree of match would be interpreted as meaning that two names do not refer to the same person (i.e. should be assigned to different person profiles). As for approach, there will likely be at least five different design cycles. One design cycle will use the clustering approach, as this approach has been used by others and will allow for comparison to prior studies [3,4,5,6]. One design cycle will use the record linkage approach, with any necessary adaptations for full-text unstructured documents. The approaches to be used for the remaining design cycles are undecided at this time, but may include decision trees, support vector machines, or even rule-based heuristics. Each design cycle will be evaluated by comparing system-produced person profiles to manually created gold standard person profiles. The design that shows the best performance will be chosen as the person resolution algorithm and implemented.

The resulting system will have a flexible architecture so that the user will be in complete control of how the ambiguity resolution proceeds. The user may decide to do no resolution at all, to do only multimorphic resolution, to do only multireferent resolution, or to do both multimorphic and multireferent resolution. The architecture will also include a weighting factor for both multimorphic resolution and multireferent resolution. The user may decide to weight both types of resolution equally, or may decide to give greater weight to either multimorphic resolution or multireferent resolution. Following implementation of the person resolution algorithm, four sets of person profiles will be produced using the entire corpus. One set will include no resolution, one set will include multimorphic resolution only, one set will include multireferent resolution only, and the final set will include both multimorphic and multireferent resolution.

Phase three involves evaluation. An intrinsic evaluation will measure the performance of the implemented person resolution algorithm by comparing the four sets of system-produced person profiles to manually created gold standard person profiles that were not used during development of the algorithm. An extrinsic evaluation will measure both the contribution of the person resolution algorithm to an information retrieval task, and the level of user satisfaction with retrieval results in the form of structured person profiles. The extrinsic evaluation will involve a second user study to collect relevance judgments and relative satisfaction. Genealogists will search for people in the collection, and will judge five sets of results for relevance of the documents returned:

- Results based on no ambiguity resolution
- Results based on automatic multimorphic resolution only
- Results based on automatic multireferent resolution only
- Results based on both automatic multimorphic and automatic multireferent resolution
- Results based on manually-created gold standard person profiles

Comparison of these relevance judgments is expected to show (1) the degree of improvement of retrieval when resolution is performed, and (2) the degree of improvement that could be expected if the person resolution algorithm performed perfectly. The genealogists will then search for other people in the collection and receive three sets of results in different formats:

- An undifferentiated ranked list of documents
- A ranked list of documents that has been differentiated as to person resolution via the use of a minimal person header containing the person's name, birth and death dates if available, and birth and death locations if available. The documents pertaining to each separate person will be grouped together under the header for that person.
- Structured person profiles

The genealogists will be asked to rank the three sets of results in terms of their relative degree of satisfaction with each format and to describe the reasons for those relative rankings. Comparison of the relative rankings and analysis of the basis for the rankings is expected to show which of the three formats is preferred most often and why.

Benefits of the study

While there have been numerous prior attempts to solve the problem of multimorphic ambiguity, both within and across documents, there have been few prior attempts to solve the problem of multireferent ambiguity. There have been no prior attempts to solve the problem of multireferent ambiguity within a single document; other studies assume one referent per discourse [3,4,5,6], but this study makes no such assumption. A person's parents, grandparents, and even great-grandparents may be mentioned in a biography, and the same names were often shared across the generations.

Limitations of the study

The limitations of this study are primarily corpus-based. The corpus used for this study is a compilation of carefully-written biographies that are not likely to contain a large number of spelling variations due to typographical or transcription errors as might be the case with more historical documents or even more current news texts. Spelling variations can be manually introduced into this corpus to help overcome this limitation. The corpus covers a set of names that primarily follow the typically American naming pattern of a first name, optional middle name, and a surname, with optional titles and optional suffixes. The person resolution algorithm would need to be adjusted to deal effectively with names from other cultures.

Conclusion

In this paper I have presented an outline of the research proposed for my dissertation. The research proposed here will commence by March of 2004 and will conclude by May of 2005.

References

1. Biographies of Notable Americans, 1904. (1997). Ancestry.com. Orem, UT: MyFamily.com, Inc.

2. Johnson, R., & Brown, J. H. (eds.). (1904). The twentieth century biographical dictionary of notable Americans. Boston, MA: The Biographical Society.

3. Mann, G. S., & Yarowsky, D. (2003). Unsupervised personal name disambiguation. Proceedings of the Seventh CoNLL Conference held at HLT-NAACL 2003 (pp. 33-40). Association for Computational Linguistics.

4. Winchester, D., & Lee, M. (2002). Cross-document co-reference of proper names. Proceedings of CLUK 2002 .

5. Winchester, D., & Lee, M. (2002). Using proper names to cluster documents. In Acquiring (and Using) Linguistic (and World) Knowledge for Information Access: Papers from the 2002 Spring Symposium (Technical Report SS-02-09) (pp. 3-8). Menlo Park, CA: American Association for Artificial Intelligence.

6. Winchester, D., & Lee, M. (2001). Clustering documents using proper names. Proceedings of the 12th Meeting of Computational Linguistics in the Netherlands (CLIN 2001) .