1. What's in a Name: Current Methods, Applications, and Evaluation in Multilingual Name Search and Matching

Sherri Condon and Keith J. Miller, MITRE

Names of people, places, and organizations have unique linguistic properties, and they typically require special treatment in automatic processes. Appropriate processing of names is essential to achieve high-quality information extraction, speech recognition, machine translation, and information management, yet most HLT applications provide limited specialized processing of names. Variation in the forms of names can make it difficult to retrieve names from data sources, to perform co-reference resolution across documents, or to associate instances of names with their representations in gazetteers and lexicons. Name matching has become critical in government contexts for checking watchlists and maintaining tax, health, and Social Security records. In commercial contexts, name matching is essential in credit, insurance, and legal applications.

This tutorial will focus on personal names, with special attention given to Arabic names, though it will be clear that much of the material applies to other languages and to names of places and organizations. Case studies will be used to illustrate problems and approaches to solutions. Arabic names illustrate many of the issues encountered in multilingual name matching, among which are complex name structures and spelling variation due to morphophonemic alternation and competing transliteration conventions.

1.1 Tutorial Outline

1. Name matching across languages, scripts, and cultures
   - Survey of problems using Arabic case study
     * Name parts and structure (titles, initials, particles, prefixes, suffixes, nicknames, tribal names)
     * Transliteration complications (segmentation, ambiguity, incompleteness, dialect variation, acoustic mismatches, competing standards)
     * Other difficulties presented by personal names
   - Survey of approaches to solutions, advantages/disadvantages of each:
     * SOUNDEX, generic string matching (Levenshtein, n-gram, Jaro-Winkler),
     * Variant generation (pattern matching, dictionaries, gazetteers),
     * Normalization (morphological analysis, rewriting, "deep" structures)
     * Intelligent-search algorithms that incorporate linguistic knowledge in selection of string-similarity measures, parameters, and lists
   - Matching across scripts
     * Methods for data acquisition
     * Transliteration
     * Phonological interlingua
2. Evaluation of Name Search and Matching Systems
   - Development of ground-truth sets
     * Human adjudication
     * Estimation techniques
   - Case study: adjudication exercises
   - Issues in establishing ground truth: different truth for different applications
   - Metrics (precision, recall, F scores, others)
   - Case study comparing matching systems for Romanized Arabic names (based on MITRE evaluation of 9 name matching products)
   - Inter-adjudicator agreement
   - Performance and other considerations

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 299–300,
New York City, June 2006.

1.2 Target Audience

This tutorial is intended for those with interest in information retrieval and entity extraction, identity resolution, Arabic computational linguistics, and related language-processing applications. As a relatively unstudied domain, name matching is a promising area for innovation and for researchers seeking new projects.

Keith J. Miller received his Ph.D. in Computational Linguistics from Georgetown University. He spent several years working on various large-scale name matching systems. His current research activities center around multicultural name matching, machine translation, embedded HLT systems, and component and system-level evaluation of systems involving HLT components.

Sherri Condon received her Ph.D. in Linguistics from the University of Texas at Austin. In addition to several years of work in multilingual name matching and cross script name matching, she is a researcher in discourse/dialogue, entity extraction, and evaluation of machine translation and dialogue systems.