A Ground Truth Dataset for Matching Culturally Diverse Romanized Person Names

Mark D. Arehart, Keith J. Miller
The MITRE Corporation
7515 Colshire Dr., McLean, VA 22102
{marehart, cwolf, keith}@mitre.org

Abstract

This paper describes the development of a ground truth dataset of culturally diverse Romanized names in which approximately 70,000 names are matched against a subset of 700. We ran the subset as queries against the complete list using several matchers, created adjudication pools, adjudicated the results, and compiled two versions of ground truth based on different sets of adjudication guidelines and methods for resolving adjudicator conflicts. The name list, drawn from publicly available sources, was manually seeded with over 1500 name variants. These names include transliteration variation, database fielding errors, segmentation differences, incomplete names, titles, initials, abbreviations, nicknames, typos, OCR errors, and truncated data. These diverse types of matches, along with the coincidental name similarities already in the list, make possible a comprehensive evaluation of name matching systems. We have used the dataset to evaluate several open source and commercial algorithms and provide some of those results.

1 Introduction

Matching multicultural Romanized names is a difficult problem due to differenes in naming practices across cultures, variation in transliteration, database fielding, and segmentation, the presence of incomplete names, titles, initials, abbreviations, and nicknames, and various forms of data corruption such as typos, OCR errors, and truncation. The type of variation encountered depends both on the linguistic origin of the name and also on the way such names are typically represented in Western databases, which generally impose a surname/given name (SN/GN) model. There are a variety of open source and commercial algorithms that purport to do fuzzy matching of names. A comprehensive evaluation requires a truthed dataset that is large, multicultural, and realistic. In this paper we describe the creation of the first version of such a database, provide evaluation results, and suggest directions for future work.

2 Data Collection

2.1 Data Sources

We collected names from two publicly available sources. The first is the Death Master File (DMF), published by the Social Security Administration, which contains the names of about 77 million deceased holders of social security numbers[1]. The data is primarily Anglo, but because it is so large, names can be found for a broad range of linguistic groups. However, the names principally conform to an Anglo name structure, as would be expected from the source. For instance an Arabic name is likely to be represented as First Middle Last or First Initial Last, e.g. *Ahmad B Husein.*

The second source is the Mémoire des hommes (MDH), published by the French government, which lists the names of about 1.3 million deceased soldiers from 20th century wars, including Indochina and North Africa, yielding many Southeast Asian and Arabic names[2]. The Arabic names typically use Francophone-influenced transliteration, e.g. *Houcine* in place of *Husein*, and contain numerous examples of multipart names. The data is noisy and includes apparent SN/GN swaps, poor SN/GN splits, and permutations.

Using a commercial name culture classification tool, 70,000 names were chosen with an approximate cultural distribution:

- 12,000 each of Anglo, Arabic, and Hispanic,
- 6,000 each of Chinese, Korean, Russian and Southwest Asian (Farsi, Afghani, and Pakistani),
- 2,000 each of French, German, Indian, Japanese, and Vietnamese.

2.2 Name Variants

Additionally we manually created 1146 variants on 404 (about 0.6%) of the base records, averaging 2.8 variants per record. These variants are spread more or less proportionally across the various cultures. The types of name variants targeted for testing can be divided into element-level variation (affecting individual name segments) and structural variation (involving more than one segment). We have broken down these types of variation into the categories below:

1) Element variations
    a) Data errors
        i) Optical Character Recognition errors
        ii) Typos
        iii) Truncations
    b) Name particles

[1] http://www.ntis.gov/products/ssa-dmf.asp

[2] http://www.memoiredeshommes.sga.defense.gouv.fr/

   i) Segmentation, e.g. *Abd Al Rahman ~ Abdal Rahman, De Los Angeles ~ Delosangeles*
   ii) Omission, e.g. of *bin* in Arabic names or *de* in Hispanic names.

c) Short forms
   i) Abbreviations, e.g. *Muhammad ~ Mhd*
   ii) Initials, e.g. *John Smith ~ J Smith*

d) Spelling variations
   i) Alternate spellings, e.g. *Jennifer ~ Jenifer*
   ii) Transliteration, e.g. *Husayn ~ Husein*

e) Nicknames and diminutives, e.g. *Robert ~ Bob*

f) Translation variants, e.g. *Joseph ~ Giuseppe*

2) Structural variations
   a) Additions/deletions, e.g. *John Smith ~ John Charles Smith*
   b) Fielding variation: division of full name into surname and given name, or swapping given name and surname
   c) Permutations, e.g. *Clara Lucia Garcia ~ Lucia Clara Garcia*
   d) Placeholders: non-name tokens like FNU, LNU, UNK
   e) Element segmentation, e.g. *Mohamed Amin ~ Mohammedamin*

Because these types of variation, which may occur singly or in combination, go beyond superficial spelling differences, we would expect searches based on generic string matching algorithms to perform relatively poorly.

2.3 Selection of Queries

Because it is infeasible to adjudicate the results of matching the entire list of 70,000 names against itself, a subset of the list was selected as queries. We chose a size of 700, approximately 1%. The queries come from two groups: the 404 “base” records, and randomly selected records.

Using the base records as queries tests a system’s ability to match all the intended variants. The randomly selected records are not expected to match as many names, since they depend on coincidental similarity. They mainly test a system’s ability to avoid false positives. High variance in the number of true matches per query was considered a desirable feature in that it resembles many real-life name matching scenarios.

Note that because the query list is a subset of the data list, each query will trivially have an exact match. Although this inflates system scores by providing low-hanging fruit for each query, it is a constant factor that will not alter system rankings. That is assuming, however, that all systems return all the exact matches. Some systems may perform parsing or normalization operations differently on query and data list names, which could potentially result in missing an exact match.

3 Ground Truth

3.1 Adjudication Pools

Ground truth for name matches was compiled by adapting the methodology of the National Institute for Standards in Technology (NIST) Text REtrieval Conference (TREC) (Voorhees and Harman, 2000; Voorhees, 2001). Because it is impossible to adjudicate every possible query list/data list name pair, only a tiny portion of which would be good matches, it is necessary to construct adjudication pools to estimate system performance in terms of recall. In order to maximize the likelihood that the pools contain all the true matches, they are generated by combining the returns of all the available algorithms using lower-than-normal matching thresholds. The algorithms include several open source and commercial-off-the-shelf (COTS) tools, described in the results section. The pools are then adjudicated according to guidelines that have been iteratively developed and refined.

3.2 Adjudication Guidelines

As has been asserted in the evaluation work in the EAGLES project, and reiterated in the follow-on related work in ISLE, it is not possible evaluate systems without considering their use context[3]. In the case of adjudicating results for our ground truth, then, the definition of a “match” versus a “non-match” cannot be determined devoid of context, but must reflect a certain use case. The scenario envisioned here is one in which a system presents name search results to an end user who has access to additional identifying attributes in order to make a decision about an overall identity match. We assume a “high risk” environment where there is a low tolerance for false negatives, and a correspondent higher tolerance for false positives. That is, the end user is willing to sift through a fair number of spurious matches to ensure that she does not miss a potentially good identity match.

We therefore developed a set of guidelines using an intentionally “loose” truth criterion, according to which two names should be considered a match despite significant variation beyond superficial spelling differences, as long as there is some plausible relationship between the names, expressed in terms of the categories of variation presented earlier. Matching record pairs in the ground truth set therefore exist along a wide continuum, from exact matches at one extreme to pairs for which the similarity is much more distant at the other. For instance, the hypothetical names below, in which the data contained in the surname field is all caps, would be considered a possible match.

a. Mohamed BIN AHMED HAMMADI
b. Haji Muhammad Hamadi AL MASRI

Figure 1: Arabic name variants.

[3] See http://www.issco.unige.ch/projects/eagles/ and http://www.issco.unige.ch/projects/eagles/ewg99/7steps.html

Note that only two of the four tokens in (1a) are matched in (1b), and two of the five tokens in (1b) are matched to (1a). Furthermore, there are no matching elements between the surname fields.

Because of the structure of Arabic names however, the apparently mismatching elements do not necessarily conflict. *Bin Ahmed* is an optional name element meaning "son of Ahmed", *Haji* is an honorific title used by someone who has made the pilgrimage to Mecca, and *Al Masri* means "the Egyptian". It is therefore possible that these two names could belong to a single person whose full name is *Haji Mohamed Bin Ahmed Hammadi Al Masri*.

Names not in the adjudication pools are assumed to be false matches for purposes of evaluation. To the extent that this is not the case, the evaluation metric will overestimate recall. However, the relative scores are still valid, so long as each algorithm is allowed to contribute equally to the adjudication pools.

3.3 Adjudication Procedure

Because adjudicating name matches is a laborious task, we developed a web-based application to facilitate the data collection and management. Users log on to the application, which presents the potential name matches for each query. The queries are presented in random order. One screenful of matches contains up to 12 name pairs, each presented in its own box with the query name shown above each data list name. The user clicks each box containing a pair she judges as a match, which highlights the box. Unselected boxes are processed as nonmatches. Once a user has completed a screen, she cannot go back and change previous answers. This is an intentional feature as we wanted users to make decisions and move through the task in a linear fashion, rather than navigating back and forth through the match pairs. As the user works her way through a queue of queries, she can log off and back in at any time. This procedure was found to be much less taxing than annotating matches presented in a text file or spreadsheet.

3.4 Compiling Ground Truth

Because different use cases will have different levels of tolerance for false positives and false negatives, in order to make our ground truth data maximally useful, we created both "loose" and "strict" versions of ground truth. With the exception of Arabic names, we used one set of adjudication guidelines that represents a middle-of-the-road view of what should match, based on the variation taxonomy presented earlier. The guidelines are not exhaustive, and we assume that judges vary in their decisions, especially on borderline cases. Therefore we have collected at least three judgments per item and have compiled different versions of ground truth according to judgment counts. The strict version consists of the items where all judges agreed on a match. The loose version consists of items where at least one judge decided on a match. This procedure was not practical for Arabic names, however, due to the relatively larger range of name variations and possible judgments, so for Arabic we have some general guidelines supplemented by explicitly loose or strict extensions. Each adjudicator followed either the strict or loose version. As long as there is at least one of each type of judge covering all the matches, then strict and loose versions of ground truth can still be compiled as the intersection and union of match judgments.

4 Evaluation Results

Shown in Table 1 are precision, recall, and F-scores for five open source and five commercial name search algorithms, according to the strict version of ground truth. The Exact search is a case-insensitive string match on full name. Exact++ is the same but with whitespace and other non-letters removed. Metaphone (Philips, 1990) is a phonetic key. Jaro-Winkler (Jaro, 1989; Winkler, 1990) and Levenshtein (Levenshtein, 1966) are both edit-based string similarity metrics. The commercial tools, which have been anonymized, employ a variety of largely proprietary algorithms.

Algorithm	**precision**	**recall**	**F**
Exact	1.00	0.24	0.39
Exact++	1.00	0.25	0.40
Metaphone	0.84	0.32	0.46
JaroWinkler	0.84	0.34	0.48
Levenshtein	0.79	0.38	0.51
Commercial 1	0.89	0.40	0.55
Commercial 2	0.75	0.46	0.57
Commercial 3	0.64	0.52	0.58
Commercial 4	0.76	0.51	0.61
Commercial 5	0.76	0.58	0.66

Table 1. Algorithm scores.

All of the algorithms except for the exact matchers and Metaphone are configurable by threshold. The results shown are for the threshold that yielded the highest F-score. The commercial tools allow for varying amounts of customization, but for purposes of this evaluation we used out-of-the-box configuration options. It is possible that performance could be improved by manipulating various parameter settings, but that is a nontrivial effort beyond the scope of this paper.

5 Conclusion

We have shown that an evaluation corpus can be developed to address the difficult, knowledge intensive problem of name matching by adapting standard Information Retrieval evaluation methods. Initial results using these methods show a wide range of performance and indicate that specialized commercial solutions outperform the generic open-source algorithms that we have tested. The general pattern among the lower-performing solutions is high precision and low recall. The higher-performing solutions, while still favoring precision, offer a more balanced tradeoff.

6 Future Work

We plan to expand the data set and analysis methods. By including additional sources of names, we will be able to create evaluation subsets with particular distributions, for example a predominantly Hispanic or Chinese test set. Other data sources will also contribute different types of noise that must be handled by the matching algorithms. We plan to expand our tagging of name pairs so that performance can be broken down according to the variant types defined in the adjudication guidelines. The work presented here, which shows global performance metrics, is the first step toward an evaluation where algorithms are rated on individual name cultures and types of variation.

References

Jaro, M. A. (1989). Advances in Record-Linkage Methodology Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association, 89 (pp. 414-420).

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Cybernetics and Control Theory, 10(8) (pp. 707-710).

Philips, L. (1990). Hanging on the Metaphone. Computer Language 7(12) (pp. 39-43).

Voorhees, E. M. (2001). The Philosophy of Information Retrieval Evaluation. Lecture Notes in Computer Science; Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems, 2406 (pp. 355-370). London, UK: Springer-Verlag.

Voorhees, E. M. and D. Harman (2000). Overview of the Eighth Text REtrieval Conference (TREC-8). In D. Harman, editor, The Eighth Text REtrieval Conference (TREC-8), Gaithersburg, MD, USA, 2000. U.S. Government Printing Office, Washington D.C.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. American Statistical Association (pp. 354-359).